Vocabulary Patterns in Free-for-all Collaborative Indexing Systems

Wolfgang Maass, Tobias Kowatsch, and Timo Münster

Hochschule Furtwangen University (HFU)
Robert-Gerwig-Platz 1, D-78120 Furtwangen, Germany
{wolfgang.maass,tobias.kowatsch,timo.muenster}@hs-furtwangen.de

Abstract. In collaborative indexing systems users generate a big amount of metadata by labelling web-based content. These labels are known as tags and form a shared vocabulary. In order to understand the characteristics of that vocabulary, we study structural patterns of these tags by implying the theory of self-organizing systems. Therefore, we utilize the graph theoretic notion to model the network of tags and their valued connections, which represent frequency rates of co-occurring tags. Empirical data is provided by the free-for-all collaborative indexing systems Delicious, Connotea and CiteULike. First, we measure the frequency distribution of co-occurring tags. Secondly, we correlate these tags towards their rank over time. Results indicate a strong relationship among a few tags as well as a notable persistence of these tags over time. Therefore, we make the educated guess that the observed collaborative indexing systems are self-organizing systems towards a shared vocabulary building. Implications on the results are the presence of semantic domains based on high frequency rates of co-occurring tags, which reflect topics of interest among the user community. When observing those semantic domains over time, that information can be used to provide a historical or trend-setting development of the community's interests, thus enhancing collaborative indexing systems in general as well as providing a new tool to develop community-based products and services at the same time.

Key words: Metadata, tagging, shared vocabulary, online community, collaborative software, self-organizing system

1 Introduction

Cooperative, distributed labelling of content in the worldwide web is called collaborative indexing or social tagging. Within a collaborative indexing system users annotate different contents e.g.: events[1], video clips[2], music[3], pictures[4],

[1] http://upcoming.org

[2] http://youtube.com

[3] http://last.fm

[4] http://flickr.com, http://espgame.org

articles and references[5], weblogs[6] or websites[7]. These collaborative indexing systems facilitate mass categorization establishing so-called folksonomies, which is a bottom up categorization made by a large user base.

A collaborative indexing system has basically two features. First, it is used for future retrieval of self-indexed content. Secondly, it provides recommendations, which are based upon the co-occurrence of highly used tags within all annotations, whereas we call one single process of annotation an indexing task.[8] The recommendations are shown to the user by committing a tag query. For instance: content tagged with *html* will be frequently tagged with *css* as well.

The data collected within an indexing task contains the name of the user, an url linking to the content, one or more tags and time-stamp information. Therefore, the data within a collaborative indexing system is basically a network of users, tags and content in a given period of time. All tags together represent the shared vocabulary of the user community. In this paper we study the structural patterns of that vocabulary, thus focusing only on the partial network of tags. Analyzing this partial network requires some constructs of the graph theory. We assume the shared vocabulary to be a self-organizing system by means of the systems theory [1]. Hence, stable patterns as well as specific correlations are determined throughout the vocabulary.

In addition, implications on these patterns are presented. To support the requirements of self-organizing systems by reducing external restrictions and forces we choose the free-for-all collaborative indexing systems Delicious, Connotea and CiteULike for empirical data extraction, where any user can index any content element. Thus, indexing rights are not restricted as identified by Marlow et al. [2].

This paper starts with related work covering collaborative indexing systems and the systems theory. Then, we hypothesize two assumptions regarding stable patterns within the vocabulary. Afterwards, we build up a model based on the graph theoretic notion, clarify the methodic approach and present the empirical data used to prove the assumptions. Subsequently, we present and discuss the results of our analysis and draw implications on them. Finally, we give an outlook on further research.

2 Related Work

A general review on collaborative indexing systems is given by Voss [3]. Mathes [4] discusses the organization of information via tags and points out that user generated metadata is of an uncontrolled nature and fundamentally chaotic compared to a controlled vocabulary. But he also mentions that collaborative index-

[5] `http://citeulike.org`, `http://connotea.org`, `http://bibsonomy.org`

[6] `http://technorati.com`

[7] `http://del.icio.us`, `http://myweb.yahoo.com`

[8] There may also exist other recommender implementations, but we focus on the co-occurrence of highly used tags because this information is freely accessible on the web.

ing systems are highly responsive to the users needs and their vocabulary by involving them into the process of organization. Vander Wal [5] distinguishes between broad and narrow folksonomies depending on the amount of users, who tag one specific content element. He also defines the difference between pure tagging and folksonomy tagging.

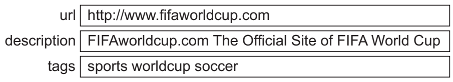
Voss [6] discovers power law distributions of tag frequency rates in Delicious and Wikipedia supporting the presence of self-organizing systems. Hotho et al. [7] and Quintarelli [8] find power law distributions according to collaborative indexing systems, too. Lund et al. [9] measure a power law distribution of user shared tags within Connotea. Results of Golder and Huberman [10] show regularities of dynamic structures within Delicious. Moreover, they introduce a classification on the semantics of tags as well as Zhichen et al. [11].

Wu et al. [12] distinguish the potential of collaborative indexing systems as a technological infrastructure for acquiring social knowledge. Millen et al. [13] study the deployment of a collaborative indexing system within a company and highlight the remarkable acceptance rate of the users as well as its personal and organizational usefulness. In addition, Damianos et al. [14], Farrell and Lau [15] as well as John and Seligmann [16] also examine the potential of collaborative indexing systems for the enterprise covering people's expertise, social networks and the integration of those systems in existing collaborative applications.

An early classification of collaborative indexing systems is done by Hammond et al. [17] confronting scholarly and general resources with links and web pages. In a more detailed classification Marlow et al. [2] distinguish the design of a system and present several user incentives. Heymann and Garcia-Molina [18] develop an algorithm, which generates a hierarchical taxonomy of a tag network. For the same purpose Mika [19] uses social network analysis on the network of users, tags and content. Hotho et al. [7] develop a search algorithm for folksonomies to find communities of interest within collaborative indexing systems. Cattuto et al. [20] design a stochastic model for the analysis of indexing tasks over time consisting of tags and users. Dubinko et al. [21] visualize tags over time with data from Flickr, whereas Zhichen et al. [11] propose an algorithm for tag suggestions to support the user within an indexing task. An overview of self-organizing systems is given by Heylighen [1].

3 Motivation

As mentioned above, this paper deals with the partial network of tags. The concept of tags is central in collaborative indexing systems. The same tags used by different users to annotate similar content show a common understanding of the users. The set of all tags utilized by the user community represents the shared vocabulary. Users and content elements are linked to each other through tags, which are also directly connected when they are used together within one indexing task. Figures 1 and 2 are representing such an indexing task as well as the resulting network of the tags *sports*, *worldcup* and *soccer.* Due to the current work, the value of those tag connections is an essential dimension, which is based on the frequency rate of tags co-occurring within all indexing tasks. A prerequisite for a measurement of this frequency is the bag-model for aggregation of tags, in which multiple tags can be assigned to one resource by multiple users as discussed by Marlow et al. [2].



url	http://www.fifaworldcup.com
description	FIFAworldcup.com The Official Site of FIFA World Cup
tags	sports worldcup soccer

Fig. 1. Graphical input mask for an indexing task

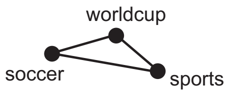


Fig. 2. Resulting network of the indexing task in Fig. 1

Prior work on stable patterns suggests that collaborative indexing systems are self-organizing systems [10, 2, 6, 8, 9]. The vocabulary - consisting of tags and generated within all indexing tasks by all users - is a part of this system, which organizes its structure by itself, without a centralized control mechanism. The users of a collaborative indexing system generate this vocabulary in a decentralized approach, not even aware of it. On its own this system evolves over time into a more stable state.

Contrary to the aforementioned work, we explore patterns emerging out of co-occurring tags. Therefore, we want to know if the power law distribution, which is common in broad folksonomies [7, 9, 8], is also applicable to the structure of co-occurring tags. This would represent a community's vocabulary, which consists of a few tags co-occurring with high frequency rates and many tags co-occurring with low frequency rates. Such a pattern - we call it tag economics - would indicate a strong consensus on a particular subpart of the community's vocabulary, from which particular interests of the users can be identified. Due to these considerations, we hypothesize the relation of co-occurring tags as follows:

H1 Let T_i be a tag and T_i^j all tags co-occurring with T_i. Then the ranked frequency distribution of all valued connections from T_i to T_i^j follows a power law curve.

Additionally, we focus on the frequency dynamics of tags over time depending on their position in the aforementioned frequency distribution. We assume that tags co-occurring with high frequency rates (higher position on the power law curve) are more stable over time than tags co-occurring with low frequency rates. This would represent persistence of the community's interests or, when tags with high frequency rates change to a low position, one can suggest a shift of the community's common understanding. Therefore, the current work has the second objective to examine the relationship of the frequency rates of co-occurring tags over time. We hypothesize this relationship as follows:

H2 The higher the frequency rates of the tags T_i^j, the more stable are they over time.

4 Model

Let an indexing task be a quadruple comprised of $< user, url, timestamp, tag^* >$. One user enters an url with none, one or more tags into a collaborative indexing system at a certain time. Only two entities are important according to our hypotheses, namely *timestamp* and *tag*. Therefore, The community's vocabulary is modelled as an undirected, valued and finite graph V within a given period of time δ. This period of time is essential, because the frequency of time based indexing tasks is subject to fluctuations, which occur in the course of a day, a week or month. Furthermore, δ can be used to affect directly the size of the vocabulary V to ease the analysis.

The vocabulary V consists of a set of nodes (here tags) and a set of valued links, which represent the frequency values of co-occurring tags. Hence, we refer to this vocabulary as the network of tags, too. The links are undirected since each tag i, which co-occurs with a tag j, also means that the tag j co-occurs with the tag i, respectively. To better handle these frequency values, the vocabulary can be described by a symmetric frequency matrix F, such that the value on the ith row and jth column represents the frequency rate of the co-occurring tags i and j over all indexing tasks within δ, denoted as $f(i, j)$. Self references are excluded since we focus only on co-occurring tags. Thus, the diagonal values $f(i, j)$ with $i = j$ are always zero. Figure 3 exemplifies an undirected, valued graph of the vocabulary V, whereas Fig. 4 shows the corresponding frequency matrix F. Based upon this graph theoretic notion and the corresponding frequency matrix, we are able to illustrate and compute the frequency distribution of co-occurring tags.

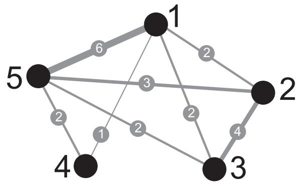
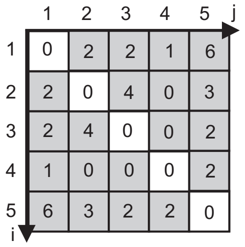


Fig. 3. Undirected, valued graph of the vocabulary V including 5 tags



i \ j	1	2	3	4	5
1	0	2	2	1	6
2	2	0	4	0	3
3	2	4	0	0	2
4	1	0	0	0	2
5	6	3	2	2	0

Fig. 4. Corresponding frequency matrix F of the vocabulary V

4.1 Method

A frequency matrix $F(\delta_1)$ is built within a given period of time. Afterwards, the frequency values $f(i, j)$ for each tag T_i are summed up. Consecutively, those cumulative frequency rates are ranked by size and confined by a limit L. This approach eliminates tags T_i with low cumulative frequency values of co-occurring tags T_i^j, because they cannot contribute any meaningful values for co-occurring tags and are therefore not relevant for further calculations. Then, N tags T_i with maximum $N_{\max}$, medium N_{med} and minimum $N_{\min}$ cumulative frequency rates are identified. Afterwards, the frequency distribution of all tags T_i^j co-occurring with each tag T_i is calculated from this selection and subsequently ranked by size. Finally, the values of these frequency distributions are normalized and utilized to conduct a curve estimation regression statistic based on the power model, whose equation is $f(r) = \beta_0 r^{\beta_1}$ with $f(r)$ estimating the frequency rate depending on the frequency rank r of a tag T_i^j. The results of the regression statistics are used to prove hypothesis 1.

The aforementioned N tags T_i are also used to prove the second hypothesis. Hence, N tags T_i with maximum $N_{\max}$, medium N_{med}, and minimum $N_{\min}$ cumulative frequency rates are identified. Afterwards, the frequency rates for each pair of co-occurring tags are written down in a time series each lasting δ_2 over I iterations. To measure the stability between co-occurring tags T_i and T_i^j, the difference D from the mean frequency of each tag co-occurrence in $N_{\max}$, N_{med}, and $N_{\min}$ is calculated over all I iterations, so they can be compared afterwards.

4.2 Empirical Data

The empirical data for the analysis was extracted from the collaborative indexing systems Delicious, Connotea and CiteULike. This information is freely accessible. Indexing rights are based on a free-for-all principle [2], thus supporting the requirements of self-organizing systems by reducing external restrictions and forces. The content is respectively of textual nature. The selected collaborative indexing systems differ in the community's size and the quantity of indexing tasks, the amount of tags, as well as the period of time in which the data was gathered. Furthermore, all indexing systems use a bag-model to aggregate tags, which is essential for our approach as mentioned in Sect. 3. Table 1 provides detailed information about the empirical data.

Table 1. Empirical data

Indexing system	Delicious	Connotea	CiteULike
Period of measurement	09/01/06 – 10/01/06	01/01/06 – 10/01/06	09/17/06 – 10/01/06
Indexing tasks (It)	452 806	92 333	3 798
It incl. at least 2 tags	269 737 (60%)	56 289 (61%)	2 430 (25%)
Tags incl. doublets	1 169 396	250 293	9 765
Distinct tags	130 776 (11%)	41 707 (17%)	3 659 (37%)
Distinct users	121 197	3 929	633
Users with at least 2 It	70 519 (58%)	2 722 (69%)	408 (64%)

5 Results

5.1 Hypothesis 1: Power Law Distribution of Co-occurring Tags

The ranked frequency distribution $f(r)$ of T_i^j tags co-occurring with a tag T_i is illustrated in Fig. 5. Thus, a power law distribution is clearly apparent in the shared vocabulary, as to be expected from a broad folksonomy like Delicious. There are many tags T_i^j with low frequency rates of co-occurring tags and few with very high frequency rates. This result is proved by the cumulative discrete co-occurrence distribution in Fig. 6, which illustrates the discrete frequency distribution. There is a remarkable gap between tags, which co-occur with only one single tag, and tags co-occurring with multiple other tags. Towards the high co-occurrence rates the curve decreases rapidly as the logarithmic scale demonstrates. Frequency rates of co-occurring tags above 100 lead to absolute frequency rates less then ten. Similar results are provided by the collaborative indexing systems Connotea and CiteULike.

The visual observations in Fig. 5 and 6 can be confirmed statistically. Therefore, Table 2 shows the median of squared reliability ($\bar{R}^2$), the median degree of freedom ($\bar{F}$) as well as the exponent $\bar{\beta}_1$ according to the power law curve estimation algorithm[9] over all corresponding tags within N_{max}, N_{med}, and N_{min}. Compared with other curve estimation algorithms, the power model performed best by far.

In particular, data from Delicious with 2007 co-occurring tags T_i^j as a median degree of freedom shows a remarkable reliability of .96 by only .01 standard deviation for N_{max}. Values with less reliability values lie nearby .80, which is still acceptable although standard deviation values show higher dynamics. Additionally, a decrease of the exponent $\bar{\beta}_1$ can be observed related to the degree of freedom by considering the data of Delicious and Connotea. This can be referred to a smoother power law curve, when less tags co-occur with a tag T_i. For this reason, the relative low degree of freedom according to CiteULike can be neglected to identify the aforementioned effect.

Due to these facts, the first hypothesis is supported by the empirical data. It is quite evident that a power law curve of co-occurring tags is obvious for tags T_i with high frequency rates, whereas the co-occurrences of middle and low ranked tags T_i show more dynamics.

5.2 Hypothesis 2: Relation between Rank and Persistence of Tags over Time

Figure 7 shows the dynamics over time of co-occurring tags against their deviations from the mean frequency. As illustrated, co-occurring tags with high frequency rates - values from $N_{\text{max}}^{T_{1-250}}$ - are more stable and have a lower scatter respectively than co-occurring tags from N_{med} or N_{min}. Figure 8 shows the

[9] Statistical software used for curve estimation: SPSS, Version 15.0.1, SPSS Inc. Chicago, USA

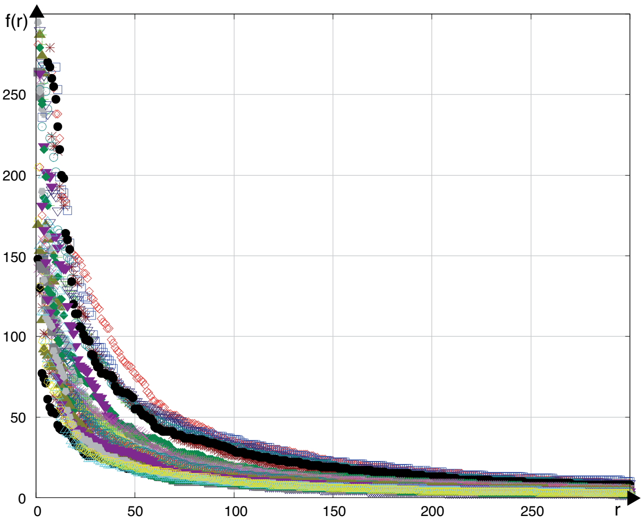


Fig. 5. Extract from the ranked frequency rates $f(r)$ of T_i^j tags co-occurring with 30 tags T_i (different symbols) based on Delicious, δ_1: 09/10/06 – 09/19/06

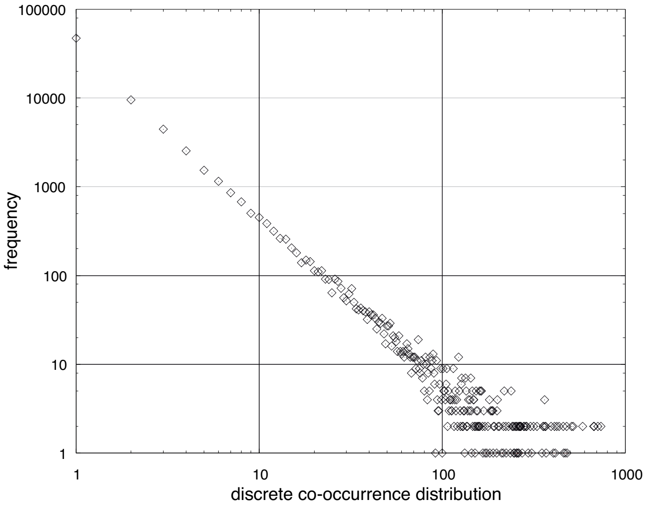


Fig. 6. Discrete co-occurrence distribution of tags T_i and T_i^j from Fig. 5 and the corresponding frequency values

Table 2. Curve estimation regression statistics based on the power law $f(r) = \beta_0 r^{\beta_1}$, $\bar{R}^2$: median of squared reliability, $\bar{F}$: median degree of freedom, standard deviations are provided in brackets

Indexing system	Delicious	Connotea	CiteULike
N / L	30 / 25	30 / 25	20 / 10
δ_1	09/10/06 – 09/19/06	01/01/06 – 09/25/06	09/15/06 – 09/25/06
$\bar{R}^2$ / $\bar{F}$ / $\bar{\beta}_1$			
for N_{max}	.96 (.01) / 2007 / -.96 (.08)	.93 (.10) / 638 / -.82 (.28)	.81 (.10) / 58 / -.46 (.18)
for N_{med}	.82 (.08) / 152 / -.51 (.12)	.86 (.06) / 88 / -.58 (.29)	.79 (.12) / 32 / -.47 (.32)
for N_{min}	.78 (.11) / 73 / -.42 (.21)	.84 (.11) / 55 / -.53 (.28)	.82 (.20) / 20 / -.65 (.40)

relative frequencies of two co-occurring tags from N_{max} and N_{med} in contrast to each other over 30 iterations with $\delta_2 = 1\,\text{day}$. This figure illustrates that the interval of N_{med} shows higher variations than the interval of N_{max}.

The basic data from all examined collaborative indexing systems with the average deviation of the mean frequency values $\bar{D}$ over N_{max}, N_{med}, and N_{min} is shown in Table 3. As a result, δ_2 and the number of indexing tasks within δ_2 are affecting the dynamics. Those in Connotea and CiteULike are much lower than the dynamics in Delicious. This often causes co-occurring tags to appear only in a small degree over all iterations and therefore, they stabilize only on a low level with low frequency rates. Thus, a small deviation from the average frequency rate over all iterations is not always indicating a high position of co-occurring tags on the power law curve. There are also situations, where tags stabilize on low frequency rates. The stability alone is therefore not a sufficient criterion for the occurrence of high frequencies. In fact, the absolute frequency rates must be observed for a positioning on the power law curve besides the deviation.

As a result, correlations between high frequency rates and their persistence over time can be concluded, but not vice versa. A change of that persistence is therefore only significant for a shared vocabulary, if the deviation of the average value appears on high frequency rates. Nevertheless, the second hypothesis is also supported by the figures of Table 3, although with less explanatory power compared to the findings of hypothesis 1.

Table 3. Deviation over the average frequency

Indexing system	Delicious	Connotea	CiteULike
N / L	250 / 30	30 / 10	15 / 5
δ_2	1 day, 09/01/06 – 09/30/06	1 month, 01/01/06 – 08/31/06	1 day, 09/16/06 – 09/30/06
$\bar{D}(N_{\text{max}})$	14.1%	22.2%	15.1%
$\bar{D}(N_{\text{med}})$	19.4%	24.0%	16.1%
$\bar{D}(N_{\text{min}})$	19.8%	27.0%	16.7%

5.3 Implications

As shown in section 5.1, we observe a frequency distribution of co-occurring tags which follows a power law curve. The examined collaborative indexing systems

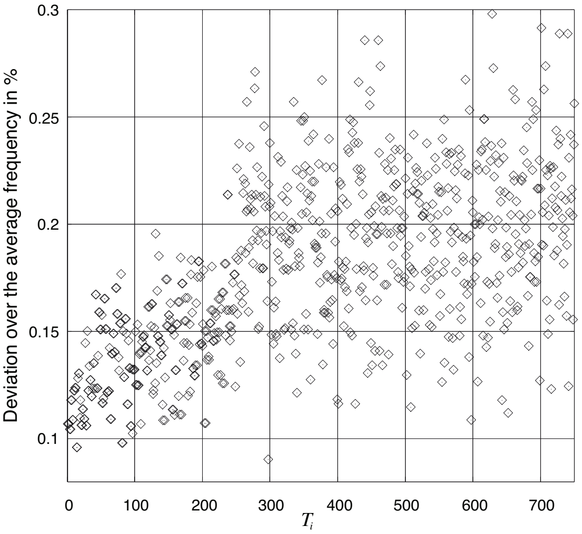


Fig. 7. Deviation over the average frequency in % of 750 tags T_i and all related tags T_i^j for $N_{\max}^{T_{1-250}}$, $N_{\text{med}}^{T_{251-500}}$ and $N_{\min}^{T_{501-750}}$ based on Delicious, between 09/01/06 – 09/30/06

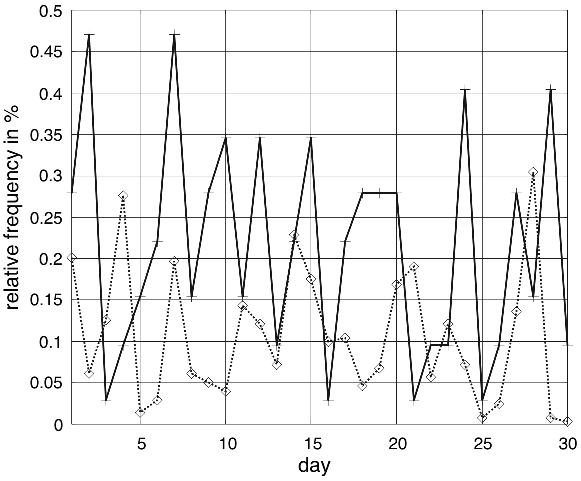


Fig. 8. Relative frequency comparison in % of 1 high (dashed) and low positioned tag T_i from Fig. 5 with $\delta_2 = 1$ day and 30 iterations based on Delicious, between 09/01/06 – 09/30/06

support a distributed approach without central control mechanisms, favoring self-organization. The observed distribution of tags means that the users have a strong consensus at least on a particular subpart of the shared vocabulary, since co-occurring tags with high frequency rates build a semantic domain. This shows some sort of tag economics within a collaborative indexing system.

A further aspect indicating a self-organizing system is the resilience of the system. Accidental errors, e.g., typos or willful sabotages of the system by users have negligible effects, because single users cannot tip the scales of a power law curve. In addition, the construct of indexing support through pre-defined tags, which is suggested to consolidate the tag usage [11, 2], would additionally support these findings by diminishing the limits of the uncontrolled vocabulary such as polysemy, synonymy/uniformity and basic level variation problems [10, 22]. Hence, we suggest higher frequency rates within the top ranked tags as well as lower rates within low ranked ones as fundamental impacts of the indexing support construct.

Another feature of a self-organizing system is the adaptation of environmental changes. In terms of a collaborative indexing system, these changes can be referred to as a shift of the community's interest, which is likewise reflected in a structural change of the vocabulary. Hence, semantic domains based upon co-occurring tags with high frequency rates may change. For instance, if the position of a tag T_i^j alters over time by means of an increase or decrease of the frequency rates according to T_i, then this progress suggests a structural change within the vocabulary and vice versa. The higher the position of this tag on the power law curve, the more significant is the structural change of the vocabulary. When this dynamic information is monitored one can observe a historical or trend-setting development of the vocabulary based upon the time-stamp of selected indexing tasks. Those trend curves of the vocabulary suggest changes within the community's interest and are useful for the particular user when searching for content elements, users or tags in the time domain.

6 Conclusion and Future Work

In this paper we studied structural patterns of user generated vocabularies within the free-for-all collaborative indexing systems Delicious, Connotea and CiteULike. The theory of self-organizing systems was implied to hypothesize patterns within those vocabularies. We built up a model based on the graph theoretic notion consisting of tags and their valued connections. This was required to calculate the frequency distribution of tags that co-occur with others, as well as to correlate those tags towards their frequency rate over time.

Results indicate that only a few co-occurring tags exist with high and many with low frequency rates, thus following a power law curve. In addition to that, co-occurring tags with high frequency rates proved to be more stable over time than those with low rates. The results were also depending on the quantity of indexing tasks. For instance, the measured values of CiteULike yielded less explanatory power than the values of Delicious. Implications are drawn through the presence of semantic domains, which are based on co-occurring tags with high frequency rates and the shift of common interests among the user community, if those high rates are fundamentally changing over time. The resulting information can be used to provide a historical or trend-setting development of the vocabulary and would not only be useful for the particular user but would also support enterprises to develop products and services, which may depend on or at least involve the interests and trends of online communities.

Due to the current work, the development of algorithms for trend information and historical time series based on the frequency distribution of co-occurring tags is an interesting area for further research. A common understanding of the user community is expressed through the tag network comprised of valued links with high frequency rates. In addition, semantic domains of more than two co-occurring tags can also be identified with techniques of the social network analysis such as centrality measurements or clustering. This network alters dynamically in a self-organizing way over time suggesting new topics or events of social, academic, technical or economic nature. Defining triggers on the observed power law curve to identify those variances requires further clarification, but would be very useful by supporting users, enterprises or public organisations in upcoming decisions.

Moreover, there is still a challenge in collaborative indexing systems featuring low indexing rates within a given period of time. This applies especially for those systems deployed in companies. Therefore, it is essential to find techniques, which permit major vocabulary coherence in such minimal systems and boost the significance of the common understanding.

References

1. Heylighen, F.: The science of self-organization and adaptivity. In: The Encyclopedia of Life Support Systems, Oxford, UK, Eolss Publishers (1999)
2. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and Hypermedia, New York, ACM Press (2006) 31–40
3. Voss, J.: Tagging, folksonomy & co - renaissance of manual indexing? ArXiv Computer Science e-prints (January 2007)
4. Mathes, A.: Folksonomies - cooperative classification and communication through shared metadata. Technical report, Graduate School of Library and Information Science, University of Illinois (December 2004)
5. Vander Wal, T.: Explaining and showing broad and narrow folksonomies http://www.personalinfocloud.com/2005/02/explaining_and_.html (February 2005)
6. Voss, J.: Collaborative thesaurus tagging the wikipedia way. ArXiv Computer Science e-prints (April 2006)
7. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In Sure, Y., Domingue, J., eds.: The Semantic Web: Research and Applications. Volume 4011 of LNAI., Heidelberg, Springer (June 2006) 411–426

8. Quintarelli, E.: Folksonomies: power to the people. Incontro ISKO Italia - UniMIB Meeting (June 2005)
9. Lund, B., Hammond, T., Flack, M., Hannay, T.: Social bookmarking tools (ii) a case study - connotea. D-Lib Magazine **11**(4) (April 2005)
10. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science **32**(2) (April 2006) 198–208
11. Zhichen, X., Yun, F., Jianchang, M., Difu, S.: Towards the semantic web: Collaborative tag suggestions. In: Collaborative Web Tagging Workshop, WWW 2006, 15th International World Wide Web Conference, Edinburgh, IW3C2 (May 2006)
12. Wu, H., Zubair, M., Maly, K.: Harvesting social knowledge from folksonomies. In: HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and Hypermedia, New York, ACM Press (2006) 111–114
13. Millen, D.R., Feinberg, J., Kerr, B.: Dogear: Social bookmarking in the enterprise. In: CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, New York, ACM Press (2006) 111–120
14. Damianos, L., Griffith, J., Cuomo, D.: Onomi: Social bookmarking on a corporate intranet. In: Collaborative Web Tagging Workshop, WWW 2006, 15th International World Wide Web Conference, Edinburgh, IW3C2 (May 2006)
15. Farrell, S., Lau, T.: Fringe contacts: People-tagging for the enterprise. In: Collaborative Web Tagging Workshop, WWW 2006, 15th International World Wide Web Conference, Edinburgh, IW3C2 (May 2006)
16. John, A., Seligmann, D.: Collaborative tagging and expertise in the enterprise. In: Collaborative Web Tagging Workshop, WWW 2006, 15th International World Wide Web Conference, Edinburgh, IW3C2 (May 2006)
17. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social bookmarking tools (i): A general review. D-Lib Magazine **11**(4) (April 2005)
18. Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Computer Science Department, Stanford University (April 2006)
19. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: 4th International Semantic Web Conference (ISWC 2005). (2005)
20. Cattuto, C., Loreto, V., Pietronero, L.: Collaborative tagging and semiotic dynamics. ArXiv Computer Science e-prints (May 2006)
21. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: Visualizing tags over time. In: WWW '06: Proceedings of the 15th international conference on World Wide Web, New York, ACM Press (2006) 193–202
22. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. Communications of the ACM **30**(11) (1987) 964–971