

# On Stability, Clarity, and Co-occurrence of Self-Tagging

Aixin Sun  
School of Computer Engineering  
Nanyang Technological University  
Singapore 639798  
axsun@ntu.edu.sg

Anwitaman Datta  
School of Computer Engineering  
Nanyang Technological University  
Singapore 639798  
anwitaman@ntu.edu.sg

## ABSTRACT

Most studies on tags focus on collaborative tagging systems where each resource (e.g., article, photo) can be tagged by multiple users with multiple tags. The tag usage patterns in self-tagging systems where a resource (e.g., a blog post) can only be tagged by its owner, however, have not been well studied. From the tags assigned by bloggers to their own blog posts, we obtain interesting insights on *tag distribution stability* and *tag clarity*. We further discuss the meaning of *tag co-occurrences* in the context of blogs and argue that tag networks based on co-occurrence from self-tagging system needs to be interpreted differently than that in collaborative tagging systems.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Information filtering*; H.1.2 [Models and Principles]: User/Machine Systems—*Human information processing*

## Keywords

Tag, Blog, Self-tagging, Clarity, Co-occurrence

## 1. INTRODUCTION

Tags are widely used for information organization on the Web. Numerous efforts have been made to better understand and exploit the use of tags and their usage patterns. Most existing studies focus on collaborative tagging systems (e.g., del.icio.us and CiteULike), where a resource may be tagged by multiple users with multiple tags [5, 6, 9, 11, 12, 13, 14]. In self-tagging systems, a resource can only be tagged by its creator. For instance, most blog service providers allow tagging of a post only by its owner. Being the sole annotator, a blogger therefore does not directly interact with or is influenced by other bloggers, in contrast to collaborative tagging systems. It is therefore interesting to study the possible differences between collaborative and self-tagging systems.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WSDM'09, Barcelona, Spain

Copyright 2009 ACM 978-1-60558-390-7 ...\$5.00.

In this paper, we study the tag usage pattern among bloggers using more than 15K blogs consisting of 3.3M blog posts and 29K distinct tags. The characteristics of our corpus are underlined in Table 1 with respect to four dimensions of the taxonomy of tagging systems [10]. Discussed in [10], the different dimensions of a tagging system have potential implications on the nature and the type of the resultant tags, the role of the tags, the convergence on folksonomy, among others. Hence, our speculation was that observations made on collaborative tagging systems may not hold on self-tagging systems. Through our experiments, we have made the following observations, which help identify similarities and differences between self-tagging and collaborative tagging.

- We show that collectively the tag distribution from all blog posts converges to a *stable* stage after certain time period. This suggests that despite no direct collaboration or communication, bloggers tend to reach a certain degree of consensus on using a relatively small set of tags to annotate most of their posts.
- We define the notion of *tag clarity* to measure, for a given tag, the degree of topical cohesiveness among the tagged documents. Our experiments show that most frequently used tags are indeed topic discriminative with respect to the collection as a whole, suggesting that bloggers share common knowledge on the semantics of these tags.
- The notion of tag clarity is also used to understand the co-occurrence relationship between two tags. Our initial studies show that a large number of co-occurrences between two tags does not necessarily indicate that the two tags are semantically similar. Instead, the two tags are likely to be semantically-orthogonal. This contrasts with what has been observed in collaborative tagging systems (e.g., [7]).

The rest of the paper is organized as follows. A brief introduction to our dataset is given in Section 2. Tag stability, clarity, and co-occurrence are discussed in detail in Sections 3, 4 and 5 respectively. After reviewing the related works in Section 6 we conclude in Section 7.

## 2. DATASET

The blog URLs in our corpus were randomly sampled from the English blogs listed in BlogFlux<sup>1</sup> directory across all categories from *academic* to *zookeeping*. Posts of the sampled

<sup>1</sup><http://dir.blogflux.com/>

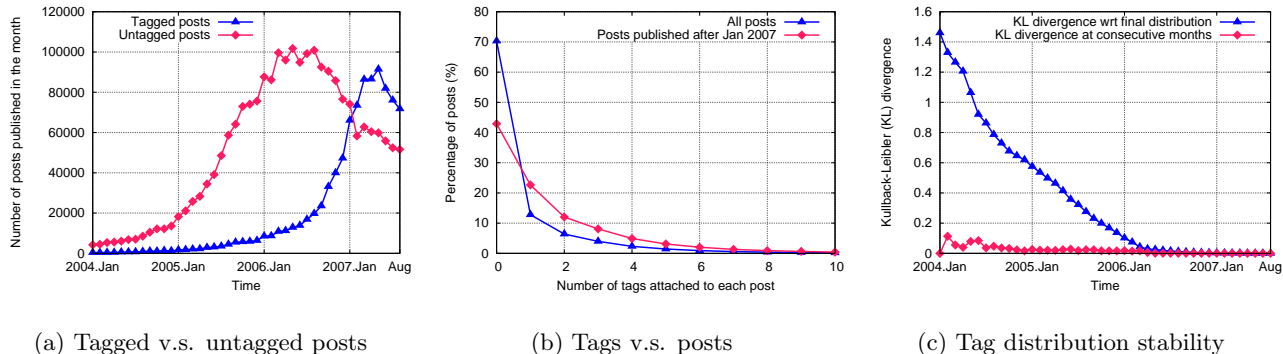


Figure 1: Tags, posts, and tag distribution

Table 1: Corpus characteristics

Dimension	Main categories
<i>Tagging rights</i>	<u>self-tagging</u> , permission-based, free-for-all
<i>Tagging support</i>	<u>blind</u> , suggested, viewable
<i>Object type</i>	<u>textual</u> , non-textual
<i>Source of material</i>	<u>user-contributed</u> , system, global

blogs hosted by `blogspot.com`<sup>2</sup> were crawled in Aug 2007. In total, 3.3M blog posts from 15,244 blogs were downloaded. Among them, nearly a million (983K) were tagged with at least one tag from 29K distinct tags.

The numbers of tagged and untagged posts published in each month from Jan 2004 to Aug 2007 are plotted in Figure 1(a). It is interesting to observe that bloggers start tagging posts much more frequently in 2006 compared to the previous two years. Among all posts published in 2007, 57% of the posts were tagged, suggesting that bloggers have adopted tagging as a way to organize blog posts, potentially realizing its utility and emulating its use from other social networking applications.

The percentages of posts with zero, one or more tags are plotted in Figure 1(b). Compared to most collaborative tagging systems where each resource is often annotated by multiple users with multiple tags, in our dataset, the chance of a blog post to be annotated by two or more tags is slim. A possible reason is that the blogger who writes the blog post has a clear understanding of the post and is able to pick up the most appropriate tag for the post (with respect to her own understanding of the semantics of all tags). The smaller number of tags assigned to blog posts leads to the question - whether co-occurrence of tags in self-tagged data carries similar functions as in collaborative tagging systems.

### 3. TAG DYNAMICS

Halpin *et al.* showed that in collaborative tagging systems the tag distribution used to collaboratively annotate a particular resource became stable after certain time period [7]. In other words, the tags that could well describe the resource are repeatedly received from multiple users. Two reasons by Golder and Huberman [6] were cited to explain the stability: *imitation of others* and *shared knowledge*. In self-tagging blogs, there is no direct interaction to influence and imitate each other, however among possibly other rea-

<sup>2</sup><http://code.google.com/apis/blogger/>

sons, bloggers may read each others’s posts and tags, which can in turn lead to shared background, leading to an implicit consensus of tag usage.

Two methods based on Kullback-Leibler (KL) divergence (see Equation 1) can be applied to test whether a distribution has converged to a stable state [7]. The first method computes the KL-divergences between distributions taken at every two consecutive time points where there is a change in the distribution. The second method computes the KL-divergences between the distribution taken at each time point and the final distribution. If the distribution reaches its stable stage, results obtained from both methods converge to and remain at zero.

$$D_{KL}(P||Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right) \quad (1)$$

Let  $T$  be the set of blog posts annotated with tag  $t$  by all bloggers and  $|T|$  be the cardinality of set  $T$ . The tag distribution by all bloggers in our setting is defined by the rank-ordered  $|T|$ ’s. We applied the two methods on tag distributions (top-5000  $|T|$ ’s<sup>3</sup>) at the end of every month and plot the KL-divergences in Figure 1(c). It shows that the tag distribution reached its stable stage in the middle of 2006, and remained stable afterwards. The final tag distribution follows a power-law distribution (not shown for the sake of space). From this we infer that bloggers do indeed develop consensus implicitly to use a relatively small set of tags to annotate most of their blog posts despite the lack of explicit interactions.

### 4. TAG CLARITY

To evaluate whether bloggers tag similar resources using similar tags (hence contributing to tag distribution stabilization), we define the notion of *tag clarity* inspired by query clarity score [4]. In simple words, the clarity score of a tag is the distance between the tag language model and the language model of the whole collection. A tag receives a high clarity score if all blog posts annotated by the tag are topically cohesive.

Equation 2 defines the unigram language model of  $t$ , denoted by  $P(w|t)$ , where  $w$  is any word and  $d$  is a document (i.e., blog post) annotated with  $t$ .  $P(w|d)$  is estimated using Jelinek-Mercer smoothing in Equation 3 with  $\lambda = 0.8$ , where  $P_{ml}(w|d)$  is the relative frequency of word  $w$  in  $d$ ,

<sup>3</sup>We follow similar setting as in [7] where the top-25  $|T|$ ’s were used.

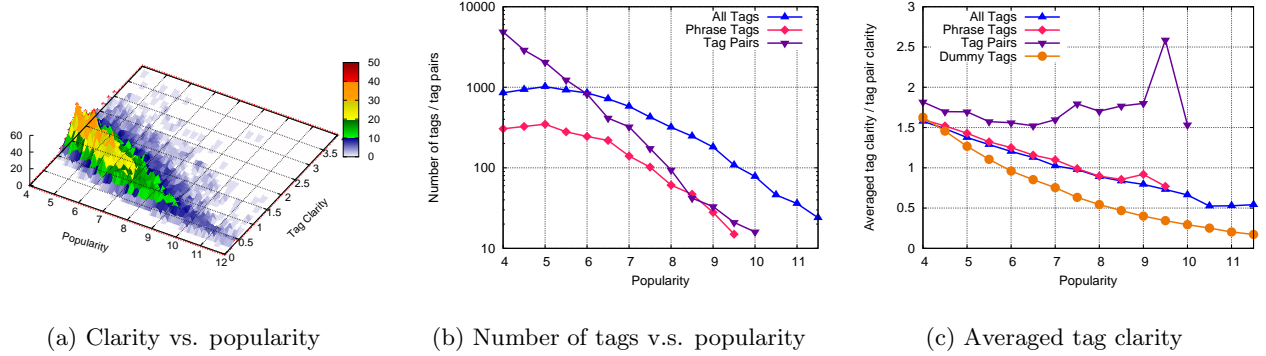


Figure 2: Clarity, tag frequency, popularity

$P(w|\mathcal{C})$  is the relative frequency of  $w$  in the whole collection  $\mathcal{C}$ . As there is no information about relative relevance of documents to a tag  $t$ , all documents are equally likely to be sampled, i.e.,  $P(d|t) = 1/|T|$ . The tag clarity score is defined in Equation 4 where  $P(w|\mathcal{C})$  denotes the unigram language model of the whole collection.

$$P(w|t) = \sum_{d \in T} P(w|d)P(d|t) \quad (2)$$

$$P(w|d) = \lambda P_{ml}(w|d) + (1 - \lambda)P(w|\mathcal{C}) \quad (3)$$

$$Clarity(t) = D_{KL}(P(w|t)||P(w|\mathcal{C})) \quad (4)$$

To avoid computational cost on less frequently used tags, we computed the clarity scores of those tags that have been used by at least 10 bloggers, and to annotate at least 16 blog posts. Among them, 2130 (or 29%) are phrase tags each consisting of at least two words and the rest are single-word tags. Figure 2(a) plots the distribution of tag clarity scores against the tag popularity where popularity of  $t$  is defined by  $\log_2(|T|)$ . Two observations can be made: (i) number of tags reduces as tag popularity increases, and (ii) clarity scores of tags decrease with the increase of popularity. The first observation holds as the tag distribution follows a power-law distribution, shown in Figure 2(b) (tag pairs will be discussed in Section 5). The second observation may be due to increase in diversity among posts with the increase in the number of blog posts annotated by a given tag.

Consider assigning a tag  $t$  to a set of documents as a sampling process of picking up documents from a large collection. If the sampling is unbiased (i.e., uniform sampling), the language model of the sampled documents  $P(w|t)$  naturally gets closer to  $P(w|\mathcal{C})$  when more documents are sampled (i.e.,  $|T|$  increases). That is, even if tags are randomly assigned to documents, in general  $Clarity(t_i) < Clarity(t_j)$  if  $|T_i| > |T_j|$ . To illustrate the impact of tag popularity, we plot the expected clarity score of dummy tags that are randomly assigned to documents in Figure 2(c). Each expected clarity score is computed using 50 dummy tags randomly assigned to blog posts with the corresponding tag popularity. Less popular tags (eg., popularity  $< 6$ ) have clarity scores close to those dummy tags while the difference between the clarity scores of popular tags and the dummy tags is significantly larger. This suggests that when a tag is frequently used, bloggers seem to reach certain level of consensus to assign the tag to some similar blog posts. Figure 2(c) also shows that the average clarity scores of phrase tags are slightly higher than that of all tags.

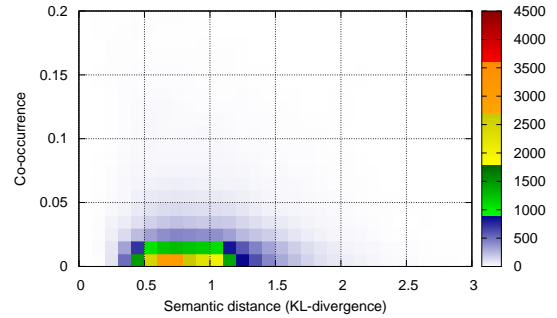


Figure 3: Semantic distance vs. Co-Occurrence

## 5. TAG CO-OCCURRENCE

Co-occurrence has been widely used to determine the relationship between two tags and to construct the network of tags for various purposes. It is often assumed that frequently co-occurred tags are similar to each other. In this section, we analyze the semantic distance between co-occurred tags.

The semantic distance (or similarity) between two tags is measured by the distance between the language models of the two tags respectively (see Equation 2). For a given tag pair  $\langle t_i, t_j \rangle$  that co-occur to annotate blog posts, their co-occurrence score is defined in Equation 5. As there is no particular order between them, the symmetric version of KL-divergence is used to measure their semantic distance, given in Equation 6. Specifically, if bloggers used to tag blog posts using semantically similar tags, then the semantic distance between the two tags should be small.

$$Cooccur(t_i, t_j) = \frac{|T_i \cap T_j|}{|T_i \cup T_j|} \quad (5)$$

$$KL(t_i, t_j) = \frac{D_{KL}(P(w|t_i)||P(w|t_j)) + D_{KL}(P(w|t_j)||P(w|t_i))}{2} \quad (6)$$

Among the 7367 tags selected in Section 4, 54,830 distinct tag pairs co-occur for at least 16 times. The number of tag pairs against the popularity ( $\log_2$ -scaled number of co-occurrences) is plotted in Figure 2(b). Similar to that of tags, the distribution follows a power-law distribution.

Figure 3 illustrates the distribution of  $KL(t_i, t_j)$ 's against  $Cooccur(t_i, t_j)$ 's. No clear pattern is observed except that most  $Cooccur(t_i, t_j)$ 's are very low. The correlation coefficient between the two scores among the 54,830 tag pairs is

0.017. That is, the co-occurrence of two tags does not suggest any semantic relationship between the two tags. However, we guess that the correlation coefficient may be affected by tags used for very broad concepts, such as life.

To further understand the co-occurrence of tags, the clarity scores of tag pairs are computed similarly in Equation 7.

$$P(w|\langle t_i, t_j \rangle) = \sum_{d \in T_i \cap T_j} P(w|d)P(d|\langle t_i, t_j \rangle) \quad (7)$$

Shown in Figure 2(c), the averaged clarity scores of tag pairs are much higher than those tags with the same popularity scores. Compared to a single tag, a tag pair could much better define the topic of a blog post. For instance,  $\langle 2008, \text{election} \rangle$ ,  $\langle \text{bush}, \text{iraq} \rangle$  and  $\langle \text{politics}, \text{news} \rangle$  are all popular tag pairs. All these tag pairs could describe blog posts more precisely than any single tag in each pair. We noticed that in our dataset many of the tag pairs are likely to be semantically-orthogonal, partially consistent with [13]. As discussed in [6], tags are more for personal use than others' benefit. As a blogger has a clear understanding about her post, it is not necessary for her to tag the post with many similar tags. Rather, she may tag post with tags from different perspectives;  $\langle \text{humor}, \text{movie} \rangle$ ,  $\langle \text{life}, \text{university} \rangle$  are some examples.

## 6. RELATED WORK

Most studies on tags are based on collaborative tagging systems and the  $\langle u, t, r \rangle$  model (i.e., user  $u$  assigns tag  $t$  to resource  $r$ ) is commonly adopted. Many studies do not consider the semantic of the resource, e.g., tag usage patterns and tag distribution stability [5, 6, 7]. Without considering the semantics of the resources, it is often assumed that two tags are likely to be similar to each other if they co-occur frequently. Our preliminary study on tags for blogs, however, does not well support this assumption. Nevertheless, our finding is consistent with [13], where a probabilistic framework was proposed to resolve tag ambiguity by suggesting semantic-orthogonal tags from those tags that co-occurred with the given set of tags, using Flickr data. For tag recommendation or prediction, the semantics of the resources are often considered [9, 11] with the implicit assumption that users would tag similar resources with similar set of tags. Our work show that such an assumption can be supported even in self-tagging systems.

The notion of tag clarity is inspired by the work on predicting query performance in ad-hoc retrieval [3, 4]. Query clarity score is proposed in [4] to evaluate the topical cohesiveness among the documents matching a given query and its effectiveness was well supported by the experiments. Carmel *et al* found that other than query clarity, the aspect coverage of documents (measured by the number of topical clusters among documents) matching the query also attributed to query performance deterioration. Evaluating tag aspect coverage in the context of blogs is part of our future work.

This work is also related to studies on blog post organization using tags. Berendt [1] showed that (i) tags have a low similarity with post body and (ii) tags together with body yielded better classification accuracy than any of them alone. Experiments on clustering blog posts based on content showed that tags are useful in grouping posts into broad categories, and frequently used tags are usually good meta-labels of a cluster of blog posts [2, 8].

## 7. CONCLUSION

In this paper, we report a preliminary study on tag usage pattern in blogs, a self-tagging system. Several interesting observations are made in this work. Besides the stability of the tag distribution derived from all bloggers, we show that bloggers are likely to annotate similar blog posts with similar set of tags through the notion of tag clarity. It is also observed that the co-occurred tags may not necessarily be semantically-similar to each other, but are likely to be semantically-orthogonal. These observations are helpful in understanding the functions of tags in blogs or other self-tagging systems and worth further investigation. Moreover, the measures proposed in the paper may also be useful in some applications (e.g., tag clarity measure for blog post clustering).

## 8. ACKNOWLEDGEMENT

This work was supported by A\*STAR Public Sector R&D, Singapore, Project Number 062 101 0031.

## 9. REFERENCES

- [1] B. Berendt and C. Hanser. Tags are not metadata, but "just more content" - to some people. In *Proc. of ICWSM'07*, Colorado, USA, 2007.
- [2] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proc. of WWW'06*, pages 625–632, Edinburgh, Scotland, 2006.
- [3] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. What makes a query difficult? In *Proc. of SIGIR'06*, pages 390–397, Seattle, Washington, USA, 2006.
- [4] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proc. of SIGIR'02*, pages 299–306, Tampere, Finland, 2002.
- [5] K. Dellschaft and S. Staab. An epistemic dynamic model for tagging systems. In *Proc. of ACM HyperText'08*, pages 71–80, Pittsburgh, PA, USA, 2008.
- [6] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.
- [7] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proc. of WWW'07*, pages 211–220, Banff, Alberta, Canada, 2007.
- [8] C. Hayes, P. Avesani, and S. Veeramachaneni. An analysis of the use of tagging in a web blog recommender system. In *Proc. of IJCAI'07*, pages 2772–2777, Hyderabad, India, 2007.
- [9] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *Proc. of SIGIR'08*, pages 531–538, Singapore, 2008.
- [10] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. of ACM HyperText'06*, pages 31–40, Odense, Denmark, 2006.
- [11] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *Proc. of SIGIR'08*, pages 515–522, Singapore, 2008.
- [12] M. N. Szomszor, I. Cantador, and H. Alani. Correlating user profiles from multiple folksonomies. In *Proc. of ACM HyperText'08*, pages 33–42, Pittsburgh, PA, USA, 2008.
- [13] K. Weinberger, M. Slaney, and R. van Zwol. Resolving tag ambiguity. In *ACM Multimedia*, Vancouver, Canada, 2008.
- [14] S. A. Yahia, M. Benedikt, L. V. S. Lakshmanan, and J. Stoyanovich. Efficient network aware search in collaborative tagging sites. *Proc. VLDB Endow.*, 1(1):710–721, 2008.