Trail Patterns in Social Tagging Systems: Role of Tags as Digital Pheromones

Thomas George Kannampallil, Wai-Tat Fu

Applied Cognitive Science Lab
Beckman Institute and Human Factors Division,
University of Illinois at Urbana-Champaign,
405 North Mathews Avenue, Urbana, IL 61801
{tgk2, wfu}@illinois.edu

Abstract. The popularity of social information systems has been driven by their ability to help users manage, organize and share online resources. Though the research exploring the use of tags is relatively new, two things are widely acknowledged in the research community: (a) tags act as a medium for social collaboration, navigation and browsing and (b) an overall stable equilibrium exists among tag patterns due to the social nature of the tagging process. But there is very little agreement on what causes these stable patterns. In this paper, we take an evolutionary perspective to understand the process of tagging to investigate whether tags act as "way finders" or digital pheromones in social tagging systems. We investigate the existence of tag trails based on a semantic similarity measure among existing tags. We found that over 50% of the resources we evaluated exhibited strong trail patterns. The implications of these patterns for the design and management of social tagging systems is discussed.

Keywords: Social Tagging Systems (STS), Stigmergy, Pheromones, Web 2.0

1 Introduction

With the widespread popularity of Web 2.0, social tagging is a common feature in several web-based systems. Social tagging systems (STS) provide the flexibility to its users for annotating, organizing and categorizing their information content on the web using tags. Tags help in supporting online search, navigation, managing content and sharing with other users who have similar interests. Examples of social systems include del.icio.us (http://del.icio.us) (URLs), Flickr (http://flickr.com) (for photos), CiteULike (www.citeulike.org) (for academic papers and books). Due to its widespread popularity and use, tagging systems contain hundreds of thousands of users, tags and resources (URLs, photos, books etc.). Users of STS are free to create tags at their free will and in most cases user activities are not moderated by administrators. In this environment, one would expect relative chaos and unstructuredness in tagging patterns. But, researchers have established that stable tagging patterns arise in social tagging systems (e.g., [3, 8]).

The stability of tagging patterns has been attributed to a variety of reasons: social imitation [8], semantic organization and clustering [11], information theory based

models [4], preferential attachment [3, 10], personal preferences [14], recency of tag use [3] and a rational model [6]. The most widely accepted idea on overall tag stability is *social imitation*. Researchers [8] have argued that tag equilibrium is achieved through users' direct imitation of tags created by other users. They suggest that users directly use the tags that are previously created. The underlying assumption on the occurrence of social imitation is that tags act as a medium of *direct* coordination activity.

We believe that there are more subtle nuances by which users appropriate tags (e.g., by creating tag synonyms, semantically similar tags, or some personalized form of tags). Thus a direct imitation model may not completely explain tag pattern equilibrium. We investigate the *effect of the social medium on the coordination practices* during tagging and the role of indirect *social coordination* in the creation of stable tagging patterns. We use the principles of stigmergy, a mechanism for explaining the indirect coordination between agents (or humans), to investigate the causes for stable patterns of tagging behavior.

Stigmergy is based on the idea that physical traces of work left by others in a medium act as the basis for future coordination activities. The idea of stigmergy was developed by Grassé [9] to describe the emergence of collective coordination activities of social insects. The concept was initially used to explain the *coordination paradox* in group activities: i.e., looking at a group of social insects (Grassé looked at nest building activities of termites), it would seem that they are cooperating in an organized manner, but looking at an individual would present the picture of independent work and not being involved in the collective activity. The explanation based on stigmergy for coordination paradox is that the collective interaction is *indirect* [15]. In other words, the agents affect the behavior of other agents through indirect communication using social artifacts in the physical environment. For example, in the case of termites, nest building material; in the case of ants, ant trails are supported by pheromones. A detailed review of stigmergy can be found in Theraulaz and Bonabeau [15].

One of the most popular examples of stigmergy is the food tracing behavior in ant-colonies. Ants in the real world wander randomly between the food source and their colony. The (initial) ants leave *pheromones* in their trail as they move from the food source to the colony (and vice versa). Other ants are more likely to follow the pheromone trail rather than a random trail, thereby reinforcing a previously existing trail. If the path to the food source is long, the pheromone trail evaporates over time. Alternatively, ants choosing shorter paths will have their pheromone trails reinforced by consistent ant-traffic and shorter path length. As a result of the pheromone evaporation in longer paths, less preferred longer routes are no longer followed by ants. But when a (random) ant finds a shorter path, other ants are likely to follow that path resulting in an overall positive feedback along that path [5]. In this example, ants coordinate their action through the indirect interaction with the physical medium and pheromones act as the medium for their coordination activities.

There are many parallels between a social tagging environment and the ant-colony described above. Users in social tagging systems are driven by their *local* goals that are driven primarily by their information needs. The *global* behavior of the users are emergent and occurs as they use the social tagging system. The global behavior is spurred by the instinctive response to traces in the medium (in this case, tags). The

trail strength is developed as more users add tags that are similar to the existing tags. This leads to later users perceiving the resource in a particular way. The comparison between an ant colony and social tagging is shown in table 1.

Table 1. A comparison between an ant food searching pattern and social tagging

	Ant Colony	Social Tagging
Local behavior	Ants searching for food	Users searching STS for their information needs
Coordinating medium	Pheromones left on trails by previous ants	Tags that are added by prior users of a resource; tags act as "digital pheromones"
Trail Strength	Shorter trails have more ant traffic leading to trail strengthening	Addition of semantically similar (or same) tags leads to a stronger strength for that tag as a descriptor for the resource
Global behavior	Ants have a coordinated shortest path trail to the food source	Coordinated overall global stable pattern across the STS

The concept of stigmergy is extremely relevant in the case of social tagging activities. We believe that social tagging systems present a medium where collective action is based on the distributed cognitive activities of a set of users. The indirect coordination practices can be considered as the interconnecting glue for distributed cognitive system, with users, resources and tags, creating a balance between individual action and social phenomena. But we know very little about how these users coordinate their tagging behavior without any direct communication. We believe that the coordination activities are *indirect* and are an effect of the tagging environment. We hypothesize that similar to the ant-colony environment, tags act as digital pheromones, creating trails for users. The digital trails are strengthened by the addition of semantically similar tags by other users. Specifically, we investigate the following research questions: (a) Do tags act as "digital pheromones" to support indirect coordination? (b) How do these trails affect the overall equilibrium in tagging patterns?

We define a *trail strength index* to evaluate the strength of a trail based on the semantic similarity between tags. Let us consider the following scenario: A user adds a tag "network" to a paper on "Facebook use". The tag "network" acts as a digital trace in the medium for future users for the paper. The addition of a new tag (or modification of an existing) that is semantically similar to the existing tag would mean that the existing trail for the tag "network" is strengthened. In other words, a new user indirectly has a general degree of collective agreement on a resource. As more tags are added, the presence of semantically similar tags strengthens the trail in a certain direction. Thus, if more tags similar to the tag "network" are added, the paper becomes perceived by future users as a paper related to networks based on its tags. Addition of tags that are not semantically related to network (e.g., a tag "food") would lead to lesser strength to the tag trail. Using data from the popular scholarly social tagging system CiteULike, we investigate the development of trails in tagging networks. We found that over 50% of the resources exhibited the strong trails, where

a digital trail was created with semantically similar tags. About 23% of the resources had weak trails phenomena, where there was little agreement among the taggers (nor were semantically similar tags) by later users. We describe the implications of strong and weak trails for growth of stable patterns and discuss its importance for the design and management of social tagging systems.

2 Process of Social Tagging

There are three main components for any tagging system: users, tags created by the users and resources (URL, books, pictures, movies etc.) for which tags are assigned.

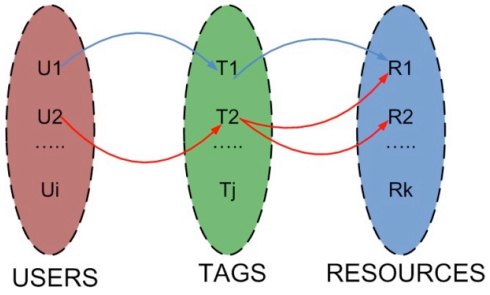

Fig 1. Process of tagging: User 1 (U1) has one tag (T1) applied to one resource (R1) (blue arrow) User 2 (U2) has one tag (T2) applied to two resources (R1 & R2) (red arrow)

The primary difference between different tagging systems is the type of resource. Resources can be different depending on the specific purpose of the social tagging system. For CiteULike (www.citeulike.com), a resource is an academic paper or book, while for del.icio.us it is a web site URL. Other media such as photos and videos have also been resources for social tagging systems. As explained earlier, a tagging system can be represented as a tuple: users, tags and resources (see U, T, and R in Figure 1). For example, consider that a user U1 applies tag T1 to a resource (R1) (see blue arrows) and user U2 applies the tag T2 to two resources R1 and R2 (see red arrows). Each of these assignments is called a tag application. Thus there are three tag applications in this system (U1, T1, R1; U2, T2, R1; U2, T2, R2). This is the widely accepted conceptual model of a social tagging system. Large social tagging systems contain hundreds of thousands of such tag applications resulting in significant interaction patterns among tags such as re-use, growth and informational value.

3 Method

In this section, the data source and the related analysis methods that were used for this study are described.

3.1 CiteULike Tagging Data

The data for this study consisted of tagging data from the popular social tagging system, CiteULike. CiteULike is an online social bookmarking service supporting the storage, sharing and organization of information on research papers. It is primarily used by researchers. CiteULike users can link papers and import references from other scholarly digital libraries. User can also add their favorite papers to their collection and assign tags to them. Our data consisted of tagging data from CiteULike over a 5-month period from November, 2004 to March, 2005. In total, there were 65,347 tag applications by 1205 users and 12067 total unique tags. The source of the data was publicly posted logs from www.citeulike.org. The data was processed to extract the user-tag-resource relationship during this time period.

3.2 Data Analysis

One of the main challenges of large datasets from public web portals is the presence of spam. While an analysis of all the tag applications is certainly useful, the results would likely be spurious due to the presence of significant spam content. As a result, we decided to process a select set of tags and manually ascertain its quality to avoid this problem. The CiteULike tagging data for the 5 month period was organized in a database. We then randomly selected 100 resources (books or papers) that had at least 5 tags and was tagged at least by 5 different users. This was done for two purposes. First, we needed to have a clear *time-sequence of tagging events* (hence the 5 tag application limit). Second, in order to establish the *effect of indirect coordination practices*, we needed to observe the effect of previous tags on the choice of future tags. The selected set of resources was manually evaluated to remove the spurious tags.

After the first 100 resources were extracted along with the corresponding users, tags and time of tags, Latent Semantic Analysis (LSA) was performed on each set [12]. LSA is used to extract and represent similarity of word meanings by comparison to large corpora of text. The LSA values reflect the general semantic similarity among words. It uses singular value decomposition, a general form of factor analysis, to condense a very large matrix of word-by-context data into a much smaller dimensional representation. LSA, as well as variations of similar statistical language techniques such as information scent [1, 2, 7, 13], had been successfully applied to explain how users interpret the relevance of link text on web pages (e.g., [1, 2]). Prior research shows significant support for the use of LSA as a method for measuring human interpretation of relatedness in text. For our analysis, LSA was performed using the algorithm available through the website at http://lsa.colorado.edu, using the general reading topic space of 300 factors.

Our analysis was conducted in the following manner: the tags created by the first user were selected. Each of these tags is compared with the rest of the tags created for that resource. In other words, the semantic similarity of each initial tag with respect to all the other available tags is computed. The matrix LSA comparison was used for our analysis. For each initial tag, we then computed the average semantic similarity score across all other available tags. We call this *trail strength index (*TSI*)*. The higher the

value of TSI, greater the semantic similarity between a tag and its follow up (trailing) tags. Lower values of trail strength indicate that successive users do not use this tag for their tagging processes.

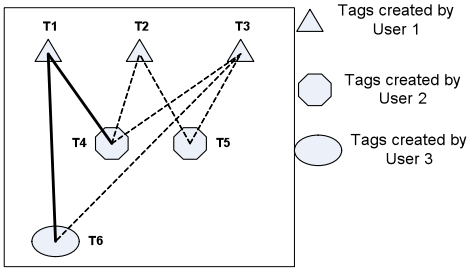

Fig 2. Evaluating the trail strength index: If a user creates semantically related tags to pre-existing tags, the strength of trail is increased (see dark line between T1 and T4, higher LSA score, therefore greater trail strength). The dotted lines show lower semantic relatedness between tags created by successive users (thus the trail strength is lower across user tagging sessions). The trail strength in this case is for tag T1 (and its semantically related tags T4 and T6) resulting in future user applying tags that are similar to T1 (or semantically similar).

The concept of strong and weak trails is shown in figure 2. As new tags are created, they are compared with prior tags to evaluate their semantic relatedness. If new tags are semantically related to previous tags, then it means users are following a trail left by previous users (hence strengthening the trail). Thus in figure 2, the semantic similarity between T1 (created by user 1) and T4 (created by user 2) is high (shown by the dark line) than with other tags. This trail is strengthened further by the addition of a new tag T6 (user 3) which is again semantically related to T1. Thus the trail of tag T1 is strengthened leading future users to perceive that T1 is the most appropriate tag for that resource. The dotted arrows show tag trails with lesser semantic relatedness. It is also likely that users may follow more than one trail. In summary, *the tags that have a higher TSI score will act as trails for that resource* resulting in future users applying that tag to the resource.

4 Results

In this section we report on the results from our analysis. As explained earlier, the analysis was conducted on 100 randomly selected resources which had more than 5 unique tags assigned to them and also had at least 5 different users tag the resource. We use examples to show two different processes that happen during tagging: strong trails and weak trails.

Of the 100 resources that we used for our analysis 53 of them fell in the strong trails category, while 23 fell in the weak trails category. The rest could not be classified into either category. That is, 76% (76/100) of the resources exhibit some form of trail properties. These trail properties were calculated based on the semantic relatedness between temporally sequenced tags. While strong trails were fairly prominent and easy to identify and explain, weakening of trails was significantly

more subtle. In the sections below we present an example of strong and weak trails based on the TSI score.

4.1 Strong Trails

Let us consider an example to explain the strong trails phenomenon. A total of 10 tags were created by 5 users (8 unique tags) for a resource (see table 2 for a sub-set of the tags). Tags were added in the order they are presented in the table (user 1 adds the tags "oscillations" and "synchronization"; then user 2 add the tag "oscillations", etc.). The trail strength index was computed by comparing each of the first two tags to the rest of the tags that were created for the resource over the five month period. Table 2 shows the computed LSA scores (and the tag trail strength index) for the two tags created by the first user.

Table 2. Average latent semantic indices for the first and second tags created by the first user for a resource are computed..

	Oscillations (User 1)	Synchronization (User 1)
Oscillations (User 2)	1	0.22
Biochemical (User 3)	0.07	0.08
Bioelectronics (User 3)	0.42	0.19
Networks (User 3)	0.04	0.23
Dynamics (User 4)	0.18	0.12
Oscillations (User 4)	1	0.22
Oscillators (User 4)	0.51	0.23
Trail Strength Index	**0.46**	**0.19**

Two important deductions can be made: (a) the first tag ("oscillations") or a highly semantically related (or similar) tag is used by future users (except User 3, who creates two tags unrelated to previous tags and one tag ("bioelectronics") which is semantically similar to pre-existing tags) leading to a fairly strong trail (avg. score =0.46, in spite of a user creating semantically unrelated tags). Thus a fairly strong "tag-trail" exists for this resource with the tag "oscillations" and (b) the second tag ("synchronization") has low semantic relatedness with other tags and it does not show decay or strengthening across all users. It is difficult to draw anything conclusive about the trail strength of the second tag.

Strong tag trails occurs by the addition of tags that are similar (or same) as the original tag. In the case of this resource, there is a stronger "tag-trail" for the tag "oscillations". Similar to the case of ants where pheromones act as a mechanism for tracing food-trail paths, strong tag trail, in terms of the semantic relatedness between a group of tags, acts as probable guidance for future users. While, it is possible to say that "oscillations" act as a digital trace (or pheromone trail) we cannot establish this unless we investigate its trail over a longer period of time. For this, we extracted all the tags that were created for this resource over the next 24 months. The next eight tags that were created for this resource were the following: oscillation, network, vibration, test, bio, frequency, oscillator, and connections. If the trail strength that has been originally established for the tag "oscillation" is strong, then more users are

likely to add a tag that is semantically related to it. For this we computed the LSA scores for the six tags that were created after the 5 month period. The mean LSA score for this set was 0.39 (the TSI score). This means that the trail of the tag "oscillator" was significantly strong and thus it is likely for other users to apply this tag (or a semantically related tag).

4.2 Weak Trails

Weakening of tag trails is the opposite of trail strengthening, i.e., the effects of tags that are created during the early phases of tagging have no effect on the more recent tags. In other words, there is no semantic relatedness between tags created earlier and recent tags. As with strong trails, we use an example to demonstrate the weakening of tag trails. As it can be noted from table 3, the initial tags that are created by user 1 (collaboration and computer-mediated) have very little semantic similarity with future tags. There does not seem to be any semantic relationship between the tags created by the first user and the ensuing tags (the TSI score for tag "collaboration" was 0.03; TSI for tag "computer-mediated" was 0.07). We also did not find any significant semantic relationships between the tags created by the second user and the third user to the later tags.

Table 3. Decaying trail strength index. This resource had more tags but some of these tags did not have LSA documentation (e.g., folksonomy, blog) so they are not shown in this list.

	Collaboration (User 1)	Computer-Mediated (User 1)
Psychology (User 2)	0.01	0.04
Social (User 2)	0.13	0.00
Email (User 2)	0.00	0.00
Technology (User 3)	0.03	0.31
Group (User 4)	0.04	0.01
Psychology (User 4)	0.01	0.04
Communication (User 5)	0.03	0.09
Trail Strength Index	**0.03**	**0.07**

Similar to strong trails, we investigated whether any of the tags created over the following 24 months had semantic similarity to these tags. Based on the set of next 8 tags, we found that there were no pairs of tags that had a tag trail index score greater than 0.1. This means that in the case of weak trails, the tags are semantically "spread" without any clear paths or trails. In such resources, the medium creates a divergence of ideas and concepts. A users' tagging choice becomes less clear in such resources.

5 Discussion

Based on the analysis of data from a social tagging system, we identified the role of tags as digital pheromones, amplifying a significant number of resources with

strong tag trails. Resources with strengthened trails exhibit clear "themes" for a resource, while the resources with weak trails had divergent tags without clear "themes" that identify the resource. While it is almost impossible to ascertain the accuracy of the themes in resources with strong trails, it is still a useful benchmark for users who have limited expertise and experience with a topic.

Based on our analysis we found the following: (a) Digital traces of tags are a function of social and semantic imitation. Tag imitation may not be direct, as described by Golder and Huberman [8], but in a more nuanced, semantic manner. (b) Stronger trails do not imply higher informational value for the tags. Tags are generally used in social navigation and search. If a resource has high tag-trail strength, it means that those tags are likely to be general descriptors of the resource and would be highly unlikely to have high informational value during search. (c) There is also the possibility that the tags that act as pheromones (with high tag-trail strength) could not be the best descriptor of the resource (possibly by the lack of or incomplete knowledge of the tag creator). This is a real possibility and could lead to spurious tags for the resource. Conversely, resources that have weak trails are likely to have tags with high informational value (i.e., able to uniquely identify a resource). Lower tag-trail strength is often a function of the nature of the tag (general vs. specific).

Additionally, it is possible to draw some interesting insights on tagging patterns based on the presence of strong and weak tag trails. The resources that have tags with strong trails are exhibit convergence of ideas and are likely to converge to a stable equilibrium much faster. This is because the tags that are added to strengthen the trail (semantically related tags) are concepts and ideas that are closely related to a central theme. One important aspect of trails which cannot be explored with a general dataset such as the one we used is to investigate the effect of expertise on the tag strength. It is likely that users' with lesser knowledge on a topic that create general tags or use their limited knowledge to create tags (resulting in semantically related tags). The creation of semantically related tags by users with lesser expertise or knowledge could also be explained based on the principle of least effort. In other words, adding existing tags (or semantically similar tags) involves less cognitive effort. Users, thereby take the easiest path of adding new tags.

In contrast, resources with weaker trails have tags that are semantically distributed or spread. It is more likely that these tags would converge at a much lower rate and are likely to be very specific (e.g., in our example, "computer-mediated"). These tags have higher informational value for search and navigation purposes.

The identification of strong and weak trails has implications for the design of social tagging systems. First, it is easy for system administrators to identify resources that have strong trails. These can be identified as resources which are less likely to be reached by users and salient tags that are not on the tag-trail can be amplified to help users in their search. Second, resources with weak trails should be amplified with tags in multiple directions to support search and retrieval. This is because these resources would otherwise not be discoverable by a large percentage of users who do not know the specific keywords.

Based on our current study, it is difficult to emphatically establish that stable equilibrium in social tagging systems is caused as a result of indirect coordination mechanisms achieved through the creation of semantically related tag-trails. But the

use of semantically tags during tagging is evident (in the case of strong trails). We need to conduct more analyses to explore whether the tags that cause tag trails contribute more towards overall equilibrium than tags that have lower tag-trail strengths. Our results show that indirect coordination using tags is a strong basis for the explanation of social tagging systems as distributed cognitive systems.

References

1. Blackmon, M.H., M. Kitajima, and P.G. Polson. Tool for Accurately Predicting Website Navigation Problems, Non-Problems, Problem Severity, and Effectiveness of Repairs. in *Proceedings of CHI 2005*. (2005). Portland, OR.
2. Brumby, D.P. and A. Howes, Strategies for Guiding Interactive Search: An Empirical Investigation into the Consequences of Label Relevance for Assessment and Selection. *Human-Computer Interaction*, (2008).
3. Cattuto, C., V. Loreto, and L. Pietronero, Semiotic Dynamics and Collaborative Tagging. *Proceedings of National Academy of Sciences*, (2007), 104, 1461-1464.
4. Chi, E.H. and T. Mytkowicz, Understanding the Efficiency of Social Tagging Systems Using Information Theory, in *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*. 2008, ACM: Pittsburgh, PA, USA.
5. Dorigo, M., V. Maniezzo, and A. Colorni, Ant System: Optimization by a Colony of Cooperating Agents. *IEEE Transactions on Systems, Man, and Cybernetics--Part B*, (1996), 26, 1, 29-41.
6. Fu, W.-T., The Microstructures of Social Tagging: A Rational Model, in *Proceedings of the ACM 2008 conference on Computer Supported Cooperative Work*. 2008, ACM: San Diego, CA, USA.
7. Fu, W.-T. and P. Pirolli, Snif-Act: A Cognitive Model of User Navigation on the World Wide Web. *Human-Computer Interaction*, (2007), 22, 355-412.
8. Golder, S.A. and B.A. Huberman, Usage Patterns of Collaborative Tagging Systems. *J. Inf. Sci.*, (2006), 32, 2, 198-208.
9. Grasse, P.P., La Reconstruction Du Nid Et Les Coordinations Interindividuelles Chez Bellicositermes Natalensis Et Cubitermes Sp. La Th´Eorie De La Stigmergie: Essai D'interpr´Etation Du Comportement Des Termites Constructeurs. *Insectes Sociaux*, (1959), 6, 41-81.
10. Halpin, H., V. Robu, and H. Shepherd, The Complex Dynamics of Collaborative Tagging, in *Proceedings of the 16th international conference on World Wide Web*. 2007, ACM: Banff, Alberta, Canada.
11. Heymann, P. and H. Garcia-Molina, Can Social Bookmarking Improve Web Search? in *First ACM International Conference on Web Search and Data Mining (WSDM'08)*. 2008.
12. Landauer, T.K. and S.T. Dumais, A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, (1997), 104, 211-240.
13. Pirolli, P. and S.K. Card, Information Foraging. *Psychological Review*, (1999), 106, 643-675.
14. Rader, E. and R. Wash, Influences on Tag Choices in Del.Icio.Us, in *Proceedings of the ACM 2008 conference on Computer Supported Cooperative Work*. 2008, ACM: San Diego, CA, USA.
15. Theraulaz, G. and E. Bonabeau, A Brief History of Stigmergy. *Artificial Life*, (1999), 5, 97-116.