An Activity-based Perspective of Collaborative Tagging

Shreeharsh Kelkar  
Avaya Labs Research  
307 Middletown-Lincroft Rd  
Lincroft, NJ  
skelkar@avaya.com

Ajita John  
Avaya Labs Research  
307 Middletown-Lincroft Rd  
Lincroft, NJ  
ajita@avaya.com

Doree Seligmann  
Avaya Labs Research  
666 5th Ave  
New York NY  
doree@avaya.com

Abstract

Collaborative tagging offers an interesting framework for studying online activity as users, topics (tags), and resources (bookmarks) get associated with each other through a folksonomy. In this paper, we consider an activity-based perspective of collaborative tagging where activity is defined as the act of associating a tag with a bookmark by a user. The perspective categorizes activities based on two defined measures: intensity and spread, which indicate the level and range, respectively, of the tagging activity, measured for both users and tags. Our block-model perspective juxtaposes two sub-perspectives: (i) A user perspective that captures the activity of users across different tags and, (ii) A tag perspective that captures the activity in tags across different users. This juxtaposition can provide an insight into different communities of users and tags. It has applications in identifying trends and types of interests in web communities as well as expertise, staffing needs and knowledge gaps in enterprise communities. Results obtained by analyzing data from a commercial tagging service offer interesting case studies.

Keywords

collaborative tagging, online activity, social network analysis, expertise, tags, bookmarks.

1. Introduction

The social aspects of the web are being felt profoundly through social networking sites such as Del.icio.us, Flickr, and YouTube [3,1,2]. These sites are revolutionizing the way content is shared and, in turn, impacting people’s access to knowledge as well as the emergence and formation of ideas, communities, and public opinion. Technologies such as collaborative tagging play a key role in enabling the sharing of content in social networks on the web. Tagging allows users to attach descriptive words or phrases to entities or bookmarks -- which may be pointers to web pages, documents, video, and audio files etc. Tags may be applied to digital representations of physical objects and people too [9]. Tagging helps to categorize resources and may enable the enhanced sharing of resources within a community of users when entities that are tagged by similar or related tags by different users are grouped together. The increasing popularity of sites such as Del.icio.us and Flickr, which rely on tagging, offers interesting aspects for studying the activity of users in tag spaces. While prior work has studied aspects such as the growth of the tag space and how individual users use tags and bookmarks [8,10,12], we develop an activity model for collaborative tagging. Activity is defined as the act of associating a tag with a bookmark by a user. The number of bookmarks associated with a certain tag indicates a certain level of activity. Our model provides a combined perspective of two views: activity by users across different tags and activity in tags across different users.

We classify users and tags into categories based on two measures: intensity and spread. Intensity is defined by the highest activity levels of a user or a tag. We use it, for instance, to determine a high-intensity user who has at least one tag with a high level of activity. Spread embodies the breadth or range of tagging activity around a certain level of activity. A high-spread high-intensity tag indicates that a wide range of users attached this tag to a large number of bookmarks. Users are classified into three intensity-based categories: high, medium and low-intensity users depending on their activity levels in one or more tags. High-intensity users are further classified into high or low-spread users depending on the number of topics (i.e. tags) in which they showed a high level of activity. Tags are categorized as high, medium or low-intensity tags depending on the activity levels in the tags by one or more users. The high-intensity tags are then further classified into high or low-spread tags depending on the number of users who showed a high level of activity in those tags.

There are two sub-perspectives in our work. These are: (i) User perspective: categories of users depicting user activity across different tags (ii) Tag perspective: categories of tags depicting activities in tags by different users. The user and tag perspectives are juxtaposed using blockmodeling approaches [5,7]. This brings to light the complex relationships between users, tags, and the tagged content. This combined perspective has interesting similarities to the notions of roles and positions in social network analysis [13].

The activity-based perspective described in this paper offers various applications in different domains. A few examples are: (1) identifying experts [11], expertise/interests, knowledge gaps and staffing needs in enterprise communities; focus groups for advertising, selecting players for games, forums and chats in web communities. (2) Understanding trends and mutations in communities by looking at the movements of users and tags through different categories (high-intensity to low-intensity, high-spread to low-spread).

Our approach has some methodological advantages: (1) it is not limited to web-pages i.e. it is independent of the type of content that is tagged. (2) It looks at the problem from a purely structural point of view i.e. no content analysis is done, although it could be augmented with content analysis; for example see [6]. (3) It is extensible. The two quantities – intensity and spread – can be granularized to the extent the data permits. Thus, there could be 5 kinds of tags/users based on intensity (rather than 3, as we have now) and 3 based on spread (rather than 2, as we have now). (4) It can be applied to collaboratively-tagged data or to automatically-tagged data or any combination of the two. For example, game playing can be characterized by automatic tags such as participants, time stamps, and points scored, in addition to user-defined collaborative tags that reflect playing experience; shopping can be characterized by automatic tags such as item.
number, cost and collaborative tags indicating user satisfaction, wants or needs. Inherent in these domains, is the notion of intensity (points scored, games played, playing experience, number of items bought, cost, user satisfaction) and spread (number of different games played, different items bought etc.) Hence, the perspective described in this paper has the potential of wide-ranging applicability.

2. Our model
In this section, we describe our activity-based model in detail. The rationale for our activity-based model is described in Section 2.1 and the categories in Section 2.2. In Section 2.3, we describe how the model can be visualized and used to obtain interesting information about tagged data, while also discussing the kinds of applications it can be used for.

Collaborative tagging allows both creators and consumers of content to generate their own keywords or phrases for annotating content: images, web-pages, videos, etc [12]. Collaboratively tagged data consists of three layers of information, as shown in Fig. 1. The layers are:
1) Bookmarks: The bottom-most layer shown in Fig. 1 is the tagged content, which we shall henceforth call bookmarks. Bookmarks are the entities that are tagged. In the case of Del.icio.us, bookmarks are pointers to web-pages (URLs); of Flickr, to photographs; of YouTube, to video-files; they may point to documents, audio files, or even physical objects.
2) Tags: Tags form the middle-layer. Tags are user-defined keywords (single words or phrases), attached to the bookmarks. The tag layer connects the users of the system to the bookmarks.
3) Users: Users are people who associate bookmarks with tags.

Fig. 1: A collaborative tagging system consists of three types of components: users of the system, the tags, and the content that is tagged, henceforth called bookmarks. The bookmarks can be web-pages, documents, video or audio files or any uniquely identifiable entity

We briefly mention some points about a tagging system:

a. Nature of bookmarks: Any collaborative tagging system has the structure shown in Fig. 1. The only difference is the type of content. The only requirement is that every bookmark be uniquely identifiable.

b. Connections within a layer: Two entities in a layer are never directly connected i.e. two users in the user layer can only be linked through their common tags or their common bookmarks; two tags only though their over-lapping users and bookmarks.

c. Connections between layers: Tags link users and bookmarks together; users and bookmarks cannot be directly linked.

d. Nature of the connections: Connections between layers, when allowed, are potentially unconstrained. Thus, the same tag could be used by multiple users, who can tag multiple bookmarks with it. Different tags may be linked to the same bookmark by the same user or by multiple users.

Formally, tagged data may be characterized as a triplet. For user U, tag T and bookmark W, we have:

\[ t(U,T,W) = \begin{cases} 
1 & \text{if user } U \text{ tags bookmark } W \text{ with tag } T \\
0 & \text{otherwise} 
\end{cases} \]  

(1)

Suppose a tagging system has N_U users, N_T tags, and N_B bookmarks. To reduce the data from 3 dimensions to 2, we do the following:

\[ f(U,T) = \sum_W t(U,T,W) \]  

(2)

Thus, every unique (U, T) pair is associated with a number, which indicates the number of unique bookmarks tagged by that user U with the tag T.

The activity of a user, U, can be represented as a N_T-dimensional (row) vector U (We will use italics to differentiate between user U and vector U):

\[ U = [f(U, T_1), f(U, T_2), ..., f(U, T_{N_T})] \]  

(3)

where T_1, T_2… T_{N_T} represent the N_T tags in the tagging system.

Similarly, the activity of a tag, T, can be represented as a N_U-dimensional (column) vector T (We will use italics to differentiate between tag T and vector T):

\[ T = [f(U_1, T), f(U_2, T), ..., f(U_{N_U}, T)]^T \]  

(4)

where U_1, U_2… U_{N_U} represent the N_U users in the tagging system.

2.1 Rationale
Our model can be characterized as a model for data reduction. In a typical collaborative tagging scenario, the number of bookmarks, users and tags is typically in thousands, if not more. Our analysis reduces the number of actors (users and tags) so that we can: (a) identify patterns and (b) easily observe relationships between actors. Our model loosely relates to standard data-reduction techniques from social network analysis, where actors are classified based on their positions and relationships between actors are characterized as roles.

Positions: According to [13], a position refers to a set of individuals who are similarly embedded in networks of relations with other individuals. For instance, a large group of friends could be sub-divided into 3 cliques, such that the friends in the same clique are friendlier to one another and somewhat distant with others not in their clique. In doing this, a social network of 100 actors can be reduced to just 3 positions, each position indicating membership of a particular clique.

Roles: Once a 100-actor network is reduced to three positions, we no longer need to consider the relationship between every pair of actors. Instead we can now consider the relationships between each position to the other 2 positions. We may find that actors in different cliques are less friendly with each other, and if the intensity of friendship is too low, then the actors in question may be “enemies”: a role. A role implies a certain relationship with other actors or positions; for example, the role of a parent involves relationships with a “child”, a “teacher” etc. The role of brother-in-law is the direct product of two relationships: marriage and sibling-hood. As per [13], roles in a social network can exist at many different levels: actors, subsets of actors and the network as a whole.
Our activity-based approach attempts to reduce both our sets of actors (i.e., users and tags) to a finite number of positions (4 for users, and 4 for tags), based on their activity. We then try and find the relationships between the different user and tag positions. Our analysis will involve the following:

1. Based on the activity of a user U, i.e., vector $U$, we classify the user U as belonging to a certain category (position). We do the same for every tag $T$ (i.e., every vector $T$).

2. We then arrange all the vectors $U$ and $T$, in a certain manner so that it brings out different relationships (roles) between the different categories of users and tags.

Finally, we examine what this activity-analysis can help us reveal about users and tags and how this can be used in different domains.

### 2.2 Activity-based categories

For all the $N_U$ users in our dataset, we attempt to classify them into 4 distinct categories. A user category defines a type of user, depending on his over-all activity represented by the row-vector $U$. Similarly, a tag category defines a type of tag, depending on its over-all activity i.e., the nature of the column vector $T$.

The number of bookmarks associated with a certain tag indicates a certain level of activity. We categorize users and tags based on two measures, intensity and spread, that relate to the activity of a user or the activity in a tag. Intensity is defined by the highest activity levels of a user or a tag. Spread embodies the breadth or range of tagging activity around a certain level. For reference, see Fig. 2.

Let $b_{ij}$ denote the number of bookmarks tagged by the user $i$ with the tag $j$. Therefore, if $U_1, U_2, ..., U_{N_U}$, are the user vectors, and $T_1, T_2, ..., T_{N_T}$ are the tag vectors in our data, the vectors in equations (3) and (4) can be written as follows:

$$ U = \begin{bmatrix} b_{U_1T_1} & b_{U_1T_2} & \cdots & b_{U_1T_{N_T}} \\ b_{U_2T_1} & b_{U_2T_2} & \cdots & b_{U_2T_{N_T}} \\ \vdots & \vdots & \ddots & \vdots \\ b_{U_{N_U}T_1} & b_{U_{N_U}T_2} & \cdots & b_{U_{N_U}T_{N_T}} \end{bmatrix} $$

(5)

$$ T = \begin{bmatrix} b_{U_1T_1} & b_{U_2T_1} & \cdots & b_{U_{N_U}T_1} \\ b_{U_1T_2} & b_{U_2T_2} & \cdots & b_{U_{N_U}T_2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{U_1T_{N_T}} & b_{U_2T_{N_T}} & \cdots & b_{U_{N_U}T_{N_T}} \end{bmatrix} $$

(6)

The intensity of a user $U$ (or a tag $T$) is a scalar which is a function of the components of the vector $U$ (or $T$). Formally:

- intensity of user $U = I(U) = l(b_{U_1T_1}, b_{U_1T_2}, \ldots, b_{U_{N_U}T_{N_T}})$
- intensity of tag $T = I(T) = l(b_{U_1T_1}, b_{U_2T_1}, \ldots, b_{U_{N_U}T_1})$

(7)

For our analysis, we set the intensity to be the largest component of the vector $U$ or $T$:

$$ I(U) = \max(b_{U_1T_1}, b_{U_1T_2}, \ldots, b_{U_{N_U}T_{N_T}}) $$

$$ I(T) = \max(b_{U_1T_1}, b_{U_2T_1}, \ldots, b_{U_{N_U}T_1}) $$

(8)

In other words, the highest number of bookmarks tagged by a user for any tag characterizes that user’s intensity and the highest number of bookmarks for a tag by any user characterizes that tag’s intensity.

Why did we choose this? Suppose that there is a user $X$ who, in a week, tagged 90 bookmarks with the tag “news” but used no other tag. Consider another user $Y$, who has used more than 40 tags, including “news” but has less than 10 bookmarks for each of those tags. We want our model to bring out users like $X$, who may or may not have used many tags, but still have singularly high levels of activity around certain tags. To be able to bring out such users, we chose the intensity for a user to be her level of activity for her most-used tag since this will help us distinguish clearly between users $X$ and $Y$. $X$ is a high-intensity user, according to our definition, even though he has only used one tag, while $Y$ is not.

Finally, we examine what this activity-analysis can help us reveal about users and tags and how this can be used in different domains.
are idealized. Real plots of tagging activity for users or tags will show fluctuations even when arranged in descending order. Moreover, while it is clear from Fig. 2, which user (or tag) is high-intensity, mid-intensity or low-intensity (it depends on the highest point of the curve) it is harder to capture the spread pictorially, when we consider real-world users and tags.

We now define our user and tag categories. Based on the intensity, we classify users/tags into 3 categories: (1) high-intensity i.e. the intensity of a user/tag is high (2) medium-intensity i.e. the intensity of a user/tag is medium and, (3) low-intensity i.e. the intensity of a user/tag is low.

We also further split the high-intensity users and tags based on their spread as follows: (1) high-spread high-intensity users/tags, which have many instances of high values of activity and, (2) low-spread high-intensity users/tags which have comparatively fewer instances of high values of activity.

How do we determine whether a certain value of the intensity or spread is high or low? This is relative to the kind of data we are analyzing. For instance, if the intensity of tag “news” is 50, then it could be a high-intensity tag if the data spanned a year, but probably not if the data spanned a day. That the tag “news” is used by 25 users at least 10 times implies that it may be high-spread, if the data was gathered over an hour, but probably not if the data was gathered over a day.

To summarize, based on intensity and spread, we assign both users and tags to one of the following four categories: (a) Category 1: high-spread high-intensity, (b) Category 2: low-spread high-intensity, (c) Category 3: medium-intensity and (d) Category 4: low-intensity. The graphs for each of the four categories are shown in Fig. 2.

2.3 Visualization and Usage

In this section, we will see how visualizing the different categories of users and tags in the form of a blockmodel helps us gain more insight into the tagging data. Suppose that we stack all N1 user-vectors together.

\[
U_{block} = \begin{bmatrix}
U_1 \\
U_2 \\
\vdots \\
U_{N_1}
\end{bmatrix} = \begin{bmatrix}
b_{U_1T_1} & b_{U_1T_2} & \cdots & b_{U_1T_{N_T}} \\
b_{U_2T_1} & b_{U_2T_2} & \cdots & b_{U_2T_{N_T}} \\
\vdots & \vdots & \ddots & \vdots \\
b_{U_{N_1}T_1} & b_{U_{N_1}T_2} & \cdots & b_{U_{N_1}T_{N_T}}
\end{bmatrix} = [T_1 \ldots T_{N_T}]
\]

Notice that stacking row-vectors representing users is the same as stacking together column vectors representing tags -- see equations (5) and (6).

In the matrix, \(U_{block}\), we now arrange the users and the tags according to their categories. That is, while stacking the vectors to create \(U_{block}\), we first put all the vectors of category 1, then category 2, and so on until category 4. We do the same for tags i.e. we stack the tag vectors together (the tag vectors are column vectors), with category 1 tag vectors at the leftmost side and category 4 tag-vectors on the rightmost side.

When all the U vectors and T vectors are thus arranged in categories, our matrix looks as shown in Fig. 3. Each rectangle inside the matrix is a block, for e.g. the block representing levels of activity for category 1 users in category 1 tags, for category 1 users in category 2 tags etc.

We make some points about the values \(x_{ij}\) in Fig. 3:

- Consider a tag in category 1 (i.e. a tag that many users have used to tag many bookmarks), i.e. the blocks \(x_{11}, x_{12}, x_{21}\) and \(x_{41}\). If the high values in this tag-vector are concentrated in \(x_{11}\), then the many high-intensity users (“experts”) of this tag are also high-intensity users of many other tags. If the high values are concentrated in \(x_{21}\), then the many high-intensity users of this tag show high levels of activity in only few other tags.

- Consider a tag in category 2, i.e. the blocks \(x_{12}, x_{22}, x_{32}\) and \(x_{42}\). If the high levels of activity in this tag-vector are concentrated in \(x_{12}\), then we know that the few high-intensity users of this tag are also high-intensity users of many other tags. If the high levels of activity are concentrated in \(x_{22}\), then we see that the few high-intensity users of this tag show high levels of activity in only few other tags.

We illustrate the above points with an example. Suppose an enterprise installs a collaborative tagging system for its employees whereby employees are allowed to tag content. That is, they are allowed to tag available documents, images, reports, etc., that form a part of the knowledge base of the enterprise. Assuming that the act of tagging is a loose measure of the employee’s expertise – “I tag therefore I know” -- it is possible, with our model, to estimate an employee’s over-all expertise or even the over-all expertise of the enterprise in a certain field.

The tags in category 2 (high-intensity low-spread) are topics with few experts, while the tags in category 1 (high-intensity high-spread) are topics with many experts. The tags in category 3 (medium-intensity) are topics, which could potentially have experts in the future. Moreover if a tag in category 2 has its high levels of activity concentrated in the block \(x_{22}\), then the topic represented by that tag has only a few expert users, who are themselves low-spread high-intensity users, and whose expertise therefore is limited to few topics (probably related to the current tag). These users then become crucial to the community since they are experts in topics, that no one else has expertise in.

More interestingly, our juxtaposition scenario can help a manager of the enterprise evaluate the enterprise’s knowledge and estimate its staffing needs. Fig. 4 shows an example.
3. Algorithm

In this section, we describe our algorithm to classify users and tags into their respective categories on the basis of their intensity and spread.

Our classification algorithm has the following two steps which are described in more detail in sections 3.1 and 3.2, respectively:

1. We first distinguish between users based on their intensity. We separate the users into high-intensity users (comprising categories 1 and 2), medium-intensity users (category 3) and low-intensity users (category 4). Then, we do the same for the tags.

2. We then take the high-intensity users and split them into two sets based on their spreads i.e. we separate out the high-intensity users into category 1 (high spread) and category 2 (low spread). The process is repeated for the tags.

3.1 Classifying users/tags on the basis of intensity

Our goal here is to classify users based on their intensity (as defined in Section 2.2), into high, medium, and low-intensity users. We apply the following steps:

1. We take all the user vectors and evaluate their intensities.

2. We arrange the user vectors in descending order of their intensities.

(3) We first separate the high-intensity users from the rest (say, non-high-intensity) users.

(4) Then we classify the non-high-intensity users into medium-intensity users and low-intensity users.

The procedure for separating a set of users into two categories—steps (3) and (4) above—is the same, both in the case of separation of high-intensity users from non-high-intensity users and in the case of separating medium-intensity users from low-intensity users. We outline this procedure below.

The vectors \( U_i, U_j, \ldots U_{N_U} \) are arranged in descending order of their intensities. In other words:

\[
I(U_i) \geq I(U_j) \geq I(U_s) \geq I(U_{N_U})
\]

i.e. \( \max(U_1) \geq \max(U_2) \geq \max(U_s) \ldots \geq \max(U_{N_U}) \) (12)

By arranging the vectors in descending order of intensities, we have made them contiguous, therefore the problem becomes a separation of high-intensity users from the set of high-intensity users and the set of non-high-intensity vectors. This is represented in Equation (13).

\[
U_{block} = \begin{bmatrix}
U_1 \\
\vdots \\
U_{n-1} \\
U_{N_U}
\end{bmatrix}
\]

We thus have to find the boundary \( n \), which divides the set of vectors \( U_{block} \) into high-intensity users and non-high-intensity users, which are represented by the matrices \( U_{high-intensity} \) and \( U_{non-high-intensity} \).

To find this boundary, we look at all \( n \) from 1 to \( N_U \). For every \( n \) we find a utility function that is given by:

\[
\text{utility}(n) = (\text{avg}(U_{high-intensity}) - \text{avg}(U_{non-high-intensity}))
\]

where

\[
\text{avg}(X) = \frac{\sum X_i}{\text{length}(X)}
\]

Note that as \( n \) increases from 1 to \( N_U \), both quantities \( \text{avg}(U_{high-intensity}) \) and \( \text{avg}(U_{non-high-intensity}) \) are non-increasing, since the rows of \( U_{block} \) are arranged in descending order of their intensities. Also, since the first term is always greater than the second term, the difference is also non-increasing. The function \( \text{utility}(n) \) is a non-increasing function of \( n \). It can be plotted as shown in Fig. 5. To choose a suitable boundary \( n \), that separates the set of high-intensity users from the set of non-high-intensity users, we choose the value of \( n \) that corresponds to the “knee” of the curve. In this case, the knee of the curve is the point where the magnitude of the slope of the curve becomes less than 1.

The reason for selecting the “knee” is as follows. The categorization of a user or a tag as high-intensity, medium-intensity or low-intensity is essentially relative. This means that given a certain dataset consisting of tags, users and bookmarks, the aim is to find certain critical thresholds above which the intensity can be considered as high and below which it is considered as low. The knee of the utility function is one way of evaluating these critical thresholds.

![Fig. 4: An application to expertise detection in the enterprise. The user-tag matrix corresponding to equation (11) and Fig. 3 is shown above. The values of the matrix are represented in green color i.e. the higher the value, the brighter the green. Notice the concentration of bright greens in the regions corresponding to \( x_{11}, x_{12}, x_{13}, x_{14}, x_{21}, x_{31}, \) and \( x_{41} \). Notice how sparse the bright greens are in \( x_{22} \).](image-url)
In this section, we describe our results by applying our activity-based analysis to a set of real-world collaboratively tagged data. In section 4.1, we describe the data we analyzed. In section 4.2, we look at the trends in the data that come to light, because of the activity-based analysis.

4.1 Our data

We used the data from the social bookmarking site, www.rawsugar.com [4]. Rawsugar is a social bookmarking site similar to Del.icio.us. Users register at the site with a username and they can tag web-pages with their own tags. Rawsugar also offers a “feed” facility, whereby users can subscribe to RSS feeds. When the feed gets updated (by, say, someone posting to a blog), the corresponding web-page is indicated as having been tagged by that user.
categories, especially high-intensity high-spread (category 1) tags. Let us consider examples of tags which fell into some of these and high-intensity low-spread (category 2) tags. The results of the categorization are shown in Table 3 where we analyzed Rawsugar data spanning five months, beginning January 2006 and ending in May 2006. In Table 1, we have the total number of distinct users, tags, and URLs that were “active” activity-based analysis to the pruned data in Table 2.

In the next section, we present the results of applying our activity-based analysis to the pruned data in Table 2.

4.2 Trends in the data

The results of the categorization are shown in Table 3 where we list the number of users and tags that fall into each category. Let us consider examples of tags which fell into some of these categories, especially high-intensity high-spread (category 1) tags and high-intensity low-spread (category 2) tags.

Category 1: One would expect users to use the tag “indian” more consistently than “india” around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to March in category 1 but disappeared for all other months. The tag “indian and india” is an example of a tag that falls into category 2. Again, this seems intuitive: an idiosyncratic tag like procrastination would probably be used by only a few people. Other tags that appear consistently in the category 1 (for at least two months) are “president”, “airlines”, “trailer” and “save”. davincicode, davinci: These tags were absent in the first four months but suddenly showed up in May, coinciding with the release of the film The Da Vinci Code. The phrase “da vinci code” showed up in May as a category 1 tag – meaning that many more users had opted to use it. “davinci” however showed up as a category 2 tag (low-spread high-intensity), meaning that a small number of users had tagged a lot of bookmarks with it.

The point being made here is that one can apply a certain semantic interpretation knowing a tag and its category i.e. whether a tag (or a user) is low-spread or high-spread, low-intensity or high-intensity. Tags that are very generic will probably not be high-intensity tags. On the other hand, too-specific tags (procrastination, ppc), will tend to end up in category 2 (used by few users), even if used highly.

4.2.1 Movement between categories
By observing the movement of a tag or a user through different categories in time, one can find out different trends in a community of users.

ie7: The tag ie7 showed up in the months of February and March, in categories 2 and 1, respectively. This may be because Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to many more users and the tag moved to category 1.

evolution: The tag evolution was a category 4 tag in January and March, disappeared in February and was a category 1 tag in April, perhaps coinciding with an increased discussion of the issue around that time.
ecryption: The tag encryption moves through all categories in this 5-month period: moving from 2, to 3, to 4, then back to 2, and ending at 1. The tag classified, makes a brief appearance in March in category 1 but disappeared for all other months.

5. Conclusions and future work
The activity-based perspective presented in this paper identifies sub-communities of interest from a collaborative tagging dataset using two measures to study two interesting facets of tagging activity – intensity or the highest levels of activity and the spread around certain levels of activity. The composition of the sub-communities identified in the perspective reveal properties of the tagging community that have applications such as finding expertise and knowledge gaps in enterprise communities. The changes in the composition of these sub-communities across time reveal trends in different topics. The approach is extensible to fine-grained granularities of intensities and spreads depending on the needs of the application and the nature of the data. If the perspective is applied to popular sites such as Flickr or YouTube, it may reveal interesting aspects such as how broad or narrow are user communities that are highly interested in photographs or

<table>
<thead>
<tr>
<th>Month</th>
<th>#Users(N_u)</th>
<th>#Tags(N_t)</th>
<th>#Bookmarks(N_b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>354</td>
<td>14320</td>
<td>57008</td>
</tr>
<tr>
<td>February</td>
<td>410</td>
<td>9958</td>
<td>46333</td>
</tr>
<tr>
<td>March</td>
<td>723</td>
<td>28462</td>
<td>79242</td>
</tr>
<tr>
<td>April</td>
<td>908</td>
<td>36573</td>
<td>94820</td>
</tr>
<tr>
<td>May</td>
<td>998</td>
<td>42548</td>
<td>133606</td>
</tr>
</tbody>
</table>

Table 1: Numbers of active users, tags and bookmarks in our analyzed dataset.

<table>
<thead>
<tr>
<th>Month</th>
<th>#Users(N_u)</th>
<th>#Tags(N_t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>249</td>
<td>1259</td>
</tr>
<tr>
<td>February</td>
<td>317</td>
<td>878</td>
</tr>
<tr>
<td>March</td>
<td>540</td>
<td>1345</td>
</tr>
<tr>
<td>April</td>
<td>673</td>
<td>1312</td>
</tr>
<tr>
<td>May</td>
<td>746</td>
<td>1024</td>
</tr>
</tbody>
</table>

Table 2: Number of users and tags after pruning.

We analyzed Rawsugar data spanning five months, beginning January 2006 and ending in May 2006. In Table 1, we have the total number of distinct users, tags, and URLs that were “active” in that month.

This implies, for instance, that in the month of January, 354 users together contributed 14320 tags to annotate 57008 webpages. Since we are interested in looking at significant relationships between users and tags, we pruned out the tags that were used by less than 5% of the users. The pruned dataset is described in Table 2.

In the next section, we present the results of applying our activity-based analysis to the pruned data in Table 2.

<table>
<thead>
<tr>
<th>Month</th>
<th>Number of Users</th>
<th>Number of Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cat 1</td>
<td>Cat 2</td>
</tr>
<tr>
<td>Jan</td>
<td>67</td>
<td>66</td>
</tr>
<tr>
<td>Feb</td>
<td>62</td>
<td>13</td>
</tr>
<tr>
<td>March</td>
<td>86</td>
<td>10</td>
</tr>
<tr>
<td>April</td>
<td>102</td>
<td>49</td>
</tr>
<tr>
<td>May</td>
<td>105</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 3: The number of users and tags in each category.

procrastination: This tag showed up in two of the five months in category 2. Again, this seems intuitive: an idiosyncratic tag like procrastination would probably be used by only a few people.

indian and india: The tag “indian” appeared consistently in the category 2 for two months (April and May). The tag “india” however appeared consistently in 4 of the 5 months, mostly as a medium-intensity tag (category 3). This seems intuitive since one would expect users to use the tag “india” more consistently than the tag “indian”.

ppc: This tag, again idiosyncratic, appears twice, each time in category 2.

Other tags that appear consistently in the category 1 (for at least two months) are “president”, “airlines”, “trailer” and “save”.

davincicode, davinci: These tags were absent in the first four months but suddenly showed up in May, coinciding with the release of the film The Da Vinci Code. The phrase “da vinci code” showed up in May as a category 1 tag – meaning that many more users had opted to use it. “davinci” however showed up as a category 2 tag (low-spread high-intensity), meaning that a small number of users had tagged a lot of bookmarks with it.

4.2.1 Movement between categories
By observing the movement of a tag or a user through different categories in time, one can find out different trends in a community of users.

ie7: The tag ie7 showed up in the months of February and March, in categories 2 and 1, respectively. This may be because Microsoft released a beta version of Internet Explorer around that time. In February, only a few users had tagged many bookmarks with it. However, in March, the interest in this topic spread to many more users and the tag moved to category 1.

evolution: The tag evolution was a category 4 tag in January and March, disappeared in February and was a category 1 tag in April, perhaps coinciding with an increased discussion of the issue around that time.

ecryption: The tag encryption moves through all categories in this 5-month period: moving from 2, to 3, to 4, then back to 2, and ending at 1. The tag classified, makes a brief appearance in March in category 1 but disappeared for all other months.

5. Conclusions and future work
The activity-based perspective presented in this paper identifies sub-communities of interest from a collaborative tagging dataset using two measures to study two interesting facets of tagging activity – intensity or the highest levels of activity and the spread around certain levels of activity. The composition of the sub-communities identified in the perspective reveal properties of the tagging community that have applications such as finding expertise and knowledge gaps in enterprise communities. The changes in the composition of these sub-communities across time reveal trends in different topics. The approach is extensible to fine-grained granularities of intensities and spreads depending on the needs of the application and the nature of the data. If the perspective is applied to popular sites such as Flickr or YouTube, it may reveal interesting aspects such as how broad or narrow are user communities that are highly interested in photographs or
videos of, say, topics such as nature, family, or war. What are the other topics that the users in these communities are highly interested in? The perspective may help identify users who are contributors to rare topics.

Future work will include the exploration of more categories, with semantic and structural analysis. Additionally, we also plan to build detailed visualizations of different sub-perspectives allowing a user to navigate through the perspectives and to switch between them.

Acknowledgements
We thank Frank Smadja of Rawsugar Inc. for allowing us to use Rawsugar’s collaborative tagging dataset.

References