Discovering Trends in Collaborative Tagging Systems

Aaron Sun[1], Daniel Zeng[1], Huiqian Li[2], and Xiaolong Zheng[2]

[1] Department of Management Information Systems, University of Arizona, Tucson, Arizona
[2] The Key Lab of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China
{asun,zeng}@email.arizona.edu, {xiaolongzheng82,pluvious1981}@gmail.com

Abstract. Collaborative tagging systems (CTS) offer an interesting social computing application context for topic detection and tracking research. In this paper, we apply a statistical approach for discovering topic-specific bursts from a popular CTS - del.icio.us. This approach allows trend discovery from different components of the system such as users, tags, and resources. Based on the detected topic bursts, we perform a preliminary analysis of related burst formation patterns. Our findings indicate that users and resources contributing to the bursts can be classified into two categories: old and new, based on their past usage histories. This classification scheme leads to interesting empirical findings.

Keywords: burst, trend discovery, collaborative tagging.

1 Introduction

Being able to identify "hot" topics and emerging trends is in critical need in many application contexts (e.g., research, business, policy making). . In recent years, Collaborative Tagging Systems (CTS) [1] , as part of social computing and in paritcular application of social software, have gained significant popularity for their revolutionary ways of re-organizing information and helping form online communities. In CTS, conceptual descriptions in the form of collections of "tags" are assigned by registered users to some Web resources they have visited. In this paper, we analyze the emergence of topic bursts using data collected from a popular CTS - Del.icio.us. We first use a widely-adopted statistical technique to discover the topic-specific trends. Given the identified topic bursts, we then study their formation patterns by examining how different types of users have contributed to the dynamic formation of the trends.

This paper is organized as follows. In the next section, we briefly review previous studies on topic burst detection. In Section 3, we describe our data collection procedure. We present our topic bursts detection method in Section 4 as well as major empirical findings. In Section 5, we mainly focus on the formation patterns of the identified bursts. Section 6 concludes with a summary of our work and possible future research directions.

C.C. Yang et al. (Eds.): ISI 2008 Workshops, LNCS 5075, pp. 377–383, 2008.

2 Related Work

Analysis of temporal data has been an active topic of research for the last few years. Among various streams of related research activities, the area of Topic Detection and Tracking (TDT) [2] is concerned with discovering topically related material in textual materials. R. Swan and J. Allan proposed a χ^2 approach for extracting significant time varying features from news articles. The rapid development of Web technologies have presented many new challenges and opportunities for TDT studies Vlachos et. al. studied MSN search engine queries that arrive over time and identify bursts and semantically similar queries. E. Amitay et. al. [3] discussed using timestamps extracted from Web pages to approximate the age of the content with the primary goal of detecting significant events and trends. The underlying value of CTS as to TDT has also been noted in recent studies. S. Golder and B. Huberman [1] performed a systematic study on the structure of Del.icio.us as well as its dynamical aspects. They discussed the feasibility of discovering bursts of popularity in bookmarks and gave a simple example. A. Hotho et. al. [4] presented a PageRank-like algorithm to discover and rank the popular topics discovered in the user-tag-resource network environment of Del.icio.us.

3 Dataset

We collected data from Del.icio.us between Nov. 10 and Nov. 15, 2007 following the steps described below. We first chose a variety of topic keywords to narrow down the focus of interest. These keywords include: “game”, “movie”, “music”, and “book”, among others. For each keyword, we downloaded a complete list of Web resources that have been tagged by this keyword. We subsequently collected their individual tagging histories. Every time when a Web resource was bookmarked by someone, we are interested in such information as the user ID of the annotator, co-occurring tags, and date of tagging. At the end, we obtained a data set covering 20 categories, with 62,263,783 tagging activities captured in total. Each tagging activity is a vector consists of user/annotator ID, bookmarked URL, the tag assigned by the user, and the date the bookmark was created. Tags may be arbitrary strings. The tagging history of any resource r is recorded on a monthly basis from year 2003 to 2007, which provides sufficient data for us to capture the overall monthly trends.

4 Tag Burst Detection

4.1 Analytical Method

We characterize a tagging activity as a vector $a = \{(u, t, r, d)|u \in U, t \in T, r \in R\}$, where U represents the entire set of users, T is the set of all relevant tags (vocabulary), R is the entire collection of Web resources being annotated, and timestamp d records the date when the tagging activity occurs. By organizing

the entire set of tagging activities by month, we can readilyconstruct a data stream $ds(t) = (d_1, d_2, ..., d_i, ...d_n)$ for any tag t of interest, with data points d_i corresponding to different tagging dates.

In the next step, we use a simple statistical model, the χ^2 model model, inspired by previous work of R. Swan and J. Allan [2], to determine if the appearance of tag t in date d is significant. This model considers tag t to be generated by a random process with an unknown stationary distribution. (We are not concerned with the actual distribution here for simplicity.) In order to verify the validity of this stationary property, one can first build a 2×2 contingency table to characterize the presence and the absence of tag t. Specifically, let N denote the number of tagging activities that include tag t in month d, and $\overline{N}$ be the number of tagging activities without tag t in month d. The 2×2 contingency table then includes both measurements in month $d = d_0$ and $d < d_0$, respectively. Given this table, we can perform a χ^2 test with one degree of freedom to measure if the stationary assumption is violated. Statistically, for a χ^2 value of 7.879, there is a 0.005 probability that a feature from a stationary process would be identified as not being stationary. We thus adopt a threshold-based strategy: for any tag t under test having a χ^2 value higher than 7.879, we conclude that the hidden tag generation process has varied and therefore classify tag t as a "burst" tag of month d.

4.2 Tag Bursts

We began with an examination of tag popularities using one of the data categories - "game". The "game" category contains 1,395,453 tagging activities, in which 45,536 unique tags have been used. Among these tags, more than half of them have been only used once, and it's not surprising to observe that the tag occurrence frequency follows a long-tail distribution (Figure 1 (a)).

In order to efficiently find the burst tags, it is necessary for us to reduce the sample size before performing the χ^2 test. We preprocessed our samples based

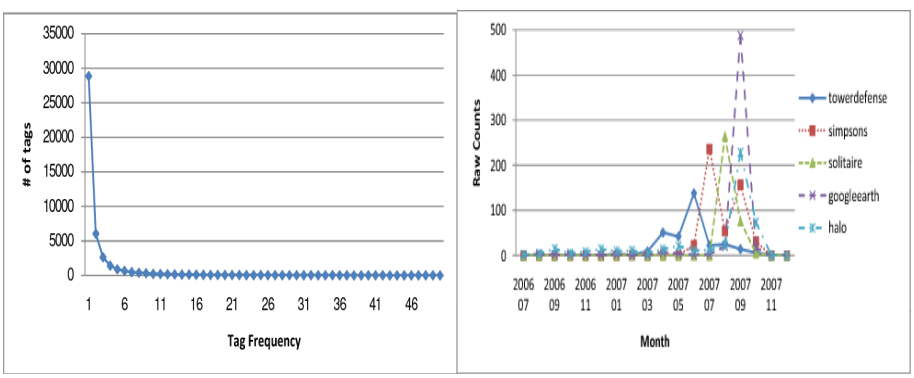

Fig. 1. (a)The distribution of tag frequency (b)Raw counts of the burst tags

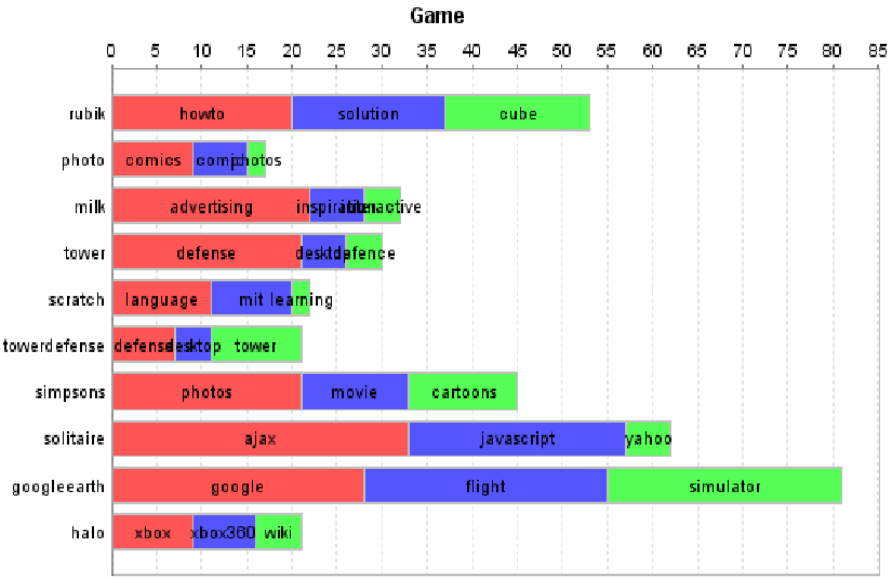

Fig. 2. The most significant tag in "game" category between Jan. to Oct., 2007

on previous studies [1] by first removing those infrequent tags whose occurrence frequencies are less than 10 per year. On the other hand, in practice, we found that some frequently occurred tags might also be reported as burst tags. For instance, in the "game" category, we have such examples as "game", "fun", and "cool", which are all top-ranked high-frequency tags. We believed that these tags had little value for us to understand the underlying tag formation and usage mechanisms. As such, the top 30 most frequently occurred tags in this category were discarded in our study.

We finally obtained a list of 1,472 candidate tags. Then starting from year 2006, we calculate the χ^2 value for each tag t in each month. If the value is above $7.879 (p < 0.005)$ we conclude that the appearance of t in that month is significant. Tag t can thus be considered as a burst tag of the month, and its χ^2 value indicates the intensity level of the burst. In Figure 1 (b), we plot the usage patterns of some detected burst tags. From the figure, it is evident that the burst periods are alike each other, having uniformly spike-shaped usage counts.

To understand the underlying events behind the identified burst tags, we performed the co-occurrence analysis aiming to study the relationship between burst formation and co-occurrence. We use a stacked bar chart in Figure 2 to illustrate our approach. The Y-axis of the chart shows the most significant tag (measured by χ^2 value) of each month from Jan. to Oct. 2007. The significance of tag t is represented by the length of the entire horizontal bar, which is equivalent to t's normalized χ^2 value. Each layer of the bar represents one of t's co-occurring pair, whose length also corresponds to its occurrence frequency. To save space,

we only draw three layers here - standing for the three most frequently co-occurred pairs. These co-occurred pairs can help us understand why some of the tags become popular. For instance, in the last week of Aug. 2007, Google released a new version of Google Earth, and this update included a fascinating hidden feature - a secret Flight Simulator - the reason that we saw Google's name extended to the area of gaming. Another example is the burst tag "halo" in Oct. 2007, which corresponded to the release of Halo 3 on Sep. 25, 2007, which is a popular Xbox 360 based game.

In the last step, we listed all the major URLs that have been annotated by the burst tags. They provide further clue about the meanings and usage of these (potentially ambiguous) tags. (Due to the page limit, we omit the URL details.)

5 Patterns of Burst Formation

Having identified a set of topic bursts, we now turn to the question of how these bursts are formed. The fundamental question we are concerned with is: when a certain tag is receiving increasing attention from users, how do these users contribute to the formation of the burst by various means? More specifically, do they create the trends by simultaneously introducing diverse information sources centered on a similar topic, or do they simply play the role of "trend-chasers" without bringing in new topics with them? In this section, we developed a simple model in an attempt to describe the browsing patterns of Del.icio.us users by making quantitative observations. We classify users and resources related to a certain tag t into two categories: *old* and *new*, based on their past usage history. For instance, *new* users U_{new} pertaining to tag t are defined as users who have not used t before date d. Likewise, *new* resources R_{new} pertain to tag t are resources having their first-time exposure to the public in d. New users turn into *old users* U_{old} when they keep on using tag t in the ensuing months. Similarly, if the *new* resources are mentioned repeatedly after d, they become *old* resources R_{old} as well. Note that the process from *new* to R_{old} is not reversible. Each user or resource can be set as *new* only once before it becomes old. We expect such a classification scheme could help us to answer the question posed above. Intuitively, if the majority of users tend to revisit their favorite topics, we will probably observe that the user population of the given burst tag contains more U_{old} than U_{new}. From the resource perspective, if we observe that the tag bursts result in higher R_{new} than R_{old}, we can conclude that it is more likely that users create the trend rather than follow it.

The classification scheme proposed above leads to stable patterns in which that bursts are dominantly contributed by new users. Empirically, we found that usually the old users only account for a small proportion of burst population (Figure 3 (a)). This stable pattern could be explained by the traditional social theory of fads/fashion [5]. In Del.icio.us, users have very little, almost zero, cost to bookmark other people's collections. The low cost of acquisition drives them to follow the preceding user's behavior blindly. For example, they are glad to bookmark those popular Web pages which are recommended by the system. As

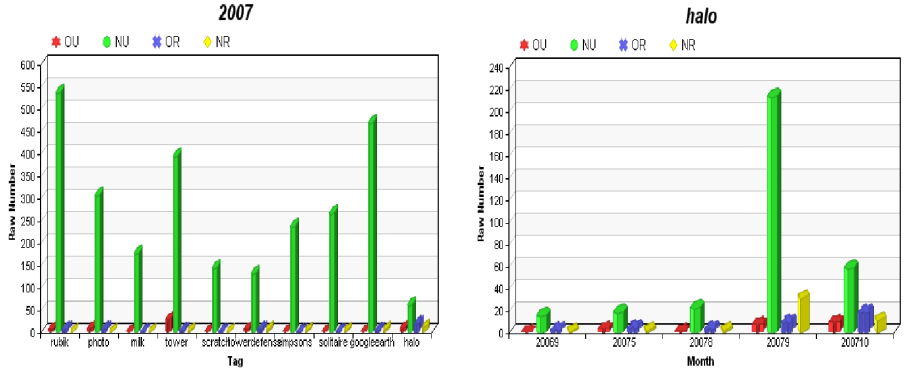

Fig. 3. User contribution to the tag bursts

is pointed out in [6], since these users do not make choices with regard to their own information, the caused mass behavior is often fragile in essence.

Another interesting finding is about the constitution of resource. As we have discussed before, the blindness of the bookmarking behavior determines that users tend to follow the trend. That can be testified by observing more R_{old} than R_{new} in the burst period (Figure 3 (a)). However, some exceptions exist. For instance, in Figure 3 (b), when Halo 3 was introduced to the public for the first time, multiple new resources were bookmarked as obviously they did not even exist before the release time. However, in the next month, the proportion of new resources shrinks to a level lower than that of the old resources.

6 Conclusions and Future Directions

CTS provide an interesting application domain for TDT study due to their large and active user base and frequent use of diverse tags. In this study, we propose to use a statistical approach to identify the topic bursts from CTS. 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 is easy to implement and extend. Given the topic bursts identified, we examine the formation patterns of the bursts.

Our future work will involve more fine-grained analysis of tag bursts (e.g., on a daily basis) In addition, we also plan to explore the impact of the community structure on the burst dynamics.

Acknowledgements

This work is supported in part by NNSFC #60621001 and #60573078, MOST #2006CB705500, #2004CB318103, and #2006AA010106, CAS #2F05N01 and #2F07C01, and NSF #IIS-0527563 and #IIS-0428241.

References

1. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (2006)
2. Swan, R., Allan, J.: Extracting significant time varying features from text. In: Proc. of the 8th Intl. Conf. on Information and knowledge management, pp. 38–45. ACM Press, New York (1999)
3. Vlachos, M., Meek, C., Vagena, Z., Gunopulos, D.: Identifying similarities, periodicities and bursts for online search queries. In: Proc. of the 2004 ACM SIGMOD Intl. Conf. on Management of Data, pp. 131–142. ACM Press, New York (2004)
4. Hotho, A., Jaschke, R., Schmitz, C., Stumme, G.: Trend detection in folksonomies. In: Proc. First International Conference on Semantics And Digital Media Technology (2006)
5. Bikhchandani, S., Hirshleifer, D., Welch, I.: Learning from the behavior of others: Conformity, fads, and informational cascades. Journal of Economic Perspectives 12(3), 151–170 (1998)
6. Bikhchandani, S., Hirshleifer, D., Welch, I.: A theory of fads, fashion, custom, and cultural change as informational cascades. The Journal of Political Economy 100(5), 992–1026 (1992)