Analysis of music tagging and listening patterns: Do tags really function as retrieval aids?

Jared Lorince[1], Kenneth Joseph[2], and Peter M. Todd[1]

[1] Cognitive Science Program
Indiana University, Bloomington, Indiana, USA
{jlorince,pmtodd}@indiana.edu
[2] Computation, Organization, and Society Program
Carnegie Mellon University, Pittsburgh, PA, USA
kjoseph@cs.cmu.edu

Abstract. In collaborative tagging systems, it is generally assumed that users assign tags to facilitate retrieval of content at a later time. There is, however, little behavioral evidence that tags actually serve this purpose. Using a large-scale dataset from the social music website Last.fm, we explore here how patterns of music tagging and subsequent listening interact in an effort to determine if there exist measurable signals of tags functioning as retrieval aids. Specifically, we describe our methods for testing if the assignment of a tag tends to lead to an increase in listening behavior. Results suggest that there exists a small but reliable effect of tags increasing listening levels, and also reveal interesting differences in which kinds of tags are most assoicated with future listening.

Keywords: Collaborative tagging, Folksonomy, Music listening, Memory cues, Retrieval aids, Personal information management

1 Introduction

In social tagging systems, users assign freeform textual labels to digital content (music, photos, web bookmarks, etc.). These individual tagging decisions are aggregated into a folksonomy [18], a "bottom-up" classificatory structure developed with little or no top-down guidance or constraints. There are a variety of reasons for which users tag content, but it is overwhelming assumed that tagging for future retrieval – assigning a tag to an item to facilitate re-finding it at a later time – is users' principal motivator. But is this a valid assumption?

Collaborative tagging systems are often designed, at least in part, as resource management platforms that expressly facilitate the use of tags as retrieval aids, but the freeform, and often social, nature of tagging opens up many other possible reasons for which a user might tag a resource. There is much non-controversial evidence for such alternative tagging motivations, (sharing resources with other users, evaluation, etc.) but the problem with the retrieval aid assumption runs deeper than there simply existing possible alternatives. There is, in fact, almost no behavioral evidence that tags are ever actually used as retrieval aids. While there is much data available user tagging habits (i.e. which terms are applied to which resources, and when), to our knowledge there is no published research providing behavioral evidence of whether or not tags, once applied to items, actually facilitate subsequent retrieval. This is an issue largely driven by a lack of data: While a web service can in principle track users' interaction with tags (for

instance, if users use tags as search terms to find tagged content), there are no available datasets containing such information, nor can it be crawled externally by researchers.

The problem is not intractable, however. While measuring how existing tags are utilized remains beyond our reach, an alternative approach is to examine how patterns of user interaction with tagged versus untagged content vary. In other words, if tags do serve as retrieval aids, we should expect users to be more likely to interact with resources (e.g. visit bookmarked pages, listen to songs, view photos, etc.) upon the application of a tag.

In the current paper we test this hypothesis using a large-scale dataset from the social music website Last.fm using, consisting of complete listening and tagging histories from more than 100,000 users. From this we extract user-artist listening time series (each representing the frequency of listening over time to a particular artist by a particular user), and compare those time series in which the user has tagged the artist, and those that are untagged. Specifically, we address the following two questions:

- RQ1: Does comparison of tagged versus untagged time series provide evidence that tagging an artist increases probability of listening to that artist in the future?
- RQ2: Do certain tags prove to be particularly associated with increases in future listening, and if so, can we identify attributes of such "retrieval-targeted" tags as opposed to others?

We describe the various analytic methods we bring to bear on these questions in Section 4, but first present related work (Section 2) and details of our dataset (Section 3). We close in Section 5 with synthesis and interpretation of our results, as well as a plan for future work.

2 Background

2.1 The formal study of folksonomies

Collaborative tagging has been considered on of the core technlogies of "Web 2.0", and has been implemented for resources as diverse as web Bookmarks (Delicious), photos (Flickr), books (LibraryThing), academic Papers (Mendley), and more. Thomas Vander Wal [18] first coined the term "folksonomy" to describe the emergent semantic structure defined by the aggregation of many individual users' tagging decisions in such a system, which of since become the target of much academic research. One of the earliest well-known and involved analyses of a collaborative tagging system is Golder and Huberman's [4] analysis of the evolution of tagging on Delicious.com, and in the same year Hotho and colleagues [8] presented a formal defintion of a folksonomy: $\mathbb{F} := (U, T, R, Y)$ [1]. U, T, and R represent, respectively, the sets of users, tags, and resources in a tagging system, while Y is aternary relation between them ($Y \subseteq U \times T \times R$). The "personomy" of a particular user (i.e. the set of resources tagged by an individual), $\mathbb{P} := (T_u, R_u, Y_u)$, can be similarly defined.

Since 2006, an extensive literature on *how* people tag has been developed, covering topics like tagging expertise [19, 20], mathematical [2] and multi-agent [11] models of tagging choices, consensus in collaborative tagging [6, 15], and

[1] This is a slight simplication. For details, See [8]

much more. Our understanding of the dynamics of tagging behavior has greatly expanded, but understadning exactly *why* people tag, on the other hand, has proven more elusive.

2.2 Why do people tag?

It is typically assumed that tags serve as retrieval aids, allowing users to re-find content to which they have applied a given tag (e.g. a user could click on search for the tag "rock" to retrieve all those songs which she has previously tagged with that term). This assumption is baked into Vander Wal's original defintion of a folksonomy, which he contends"is the result of personal free tagging of information and objects (anything with a URL) *for one's own retrieval*" [18, emphasis added]. This perspective is echoed in many studies of tagging patterns [3, 6, 4].

But while retrieval is the most commonly assumed motivation for tagging, other reasons certainly exist, and various researchers have proposed taxonomies of tagging motivation. Proposed motivational factors in tagging include personal information management (including but not limited to tagging for futrue retrieval), resource sharing, opinion expression, performance, and activism [7, 21, 1], among others. See [5] for a review.

While the development of motivational theories in tagging is useful, there is almost no work actually grounding these them in behavioral observations. The vast majority of existing work either makes inferences about motivation based on design features of a website (e.g. social motivations in tagging require that one's tags be visible to other users, [13]), employs semantic analysis and categorization of tags (e.g. the tags "to read", "classical", and "love" can all be inferred to have different uses, [21, 16]), or directly asks users why they tag using survey methods [1, 14]. The results of such approches are useful contributions to the field, but none have resulted in testable behavioral hypotheses that can confirm or refute their validity.

One notable exception is work by Körner and colleagues [9, 10, 22]. They argue that taggers can be classifed on a motivational spectrum from categorizers (who use a contrained vocabulary suitable to future browsing of taggedresources) to describers (who use a large, varied vocabulary to facilitate future keyword-based search), and have developed and tested quantifiable signals of these different motivations. The main deficiency of this approach, however, is that their hypotheses are based fully on attributes of user tag vocabularies; they present no way to test whether or not describers actually use tags, once applied, for keyword-based search and that categorizers use them for browsing.

This problem is no fault of the authors, however. Data on how users actually *use* existing tags is simply not available to researchers through any tagging system APIs (or through other methods) that we are aware of. Thus the exsiting work on tagging motivation is limited to inferring *why* people tag from *how* they tag. In presenting our methods, we are aware that they still represent an inferential approach. Our approach is distinct from those described here, however, in that we test a concrete hypothesis about how tagging should affect a behavior on which we *do* have data (interaction with tagged content, in our case music listening).

3 Dataset

Last.fm incorporates two specific features of interest to us here. First, it implements a collaborative tagging system (a "broad" folksonomy, following Vander Wal's [17] terminology, meaning that multiple users tag the same, publicly available content) in which users can label artist, albums, and songs. Second, the service tracks users' listening habits both on the website itself and on media players (e.g. iTunes) via a software plugin. This tracking process is known as "scrobbling", and each timestamped instance of a user listening to a particular song is termed a "scrobble".

Here we utilize an expanded version of a dataset described in earlier work [11, 12] that includes the full tagging histories of approximately 1.9 million Last.fm users, and full listening histories from a subset of those users (approximately 100,000) for a 90-month time window (July 2005 - December 2012, inclusive). Data were collected via a combination of the Last.fm API and direct scraping of publicly available user profile pages. For further details of the crawling process, see [11, 12].

For our current purposes, we consider only those users for which we have both tagging and listening histories. For each user, we extract one time series for each unique artist listened to by that user. Each user-artist listening time series consists of a given users's monthly listening frequency to a particular artist for each month in our data collection period, represented as a 90-element vector.

User tagging histories are only available at monthly time resolution, so we also downsample scrobble data (which is recorded to second precision) to monthly playcounts as well. Furthermore, we perform all analyses here at the level of artists, rather than individual songs. Thus every song scrobbled is treated as a listen to the corresponding artist, and all annotations (which can be applied to songs, albums, or artists) are treated as annotations of the corresponding artist. Our choice to perform all analyses at the level of artists, rather than individual songs, is based on the facts that (a) listening and tagging data for any particular song tends to be very sparse, and (b) the number of time series resulting from considering each unique song listened to by each user would be prohibitively large.

The over 2 billion individual scrobbles in our dataset define a total of ~ 95 million user-artist listening time series. In ~ 6 million of these cases, the user has assigned at least one tag to the artist (or to a song or album by that artist) within the collection period (we refer to these as tagged time series), while in the remaining cases (~ 89 million) the user has never tagged the artist. We summarize these high level dataset statistics in Table 1. Comparison of these these tagged and untagged listening time series is the heart of the analyses presented in the next section.

4 Analyses & Results

4.1 Comparison of tagged and untagged time series

Our principal research question is whether listening patterns to tagged versus untagged content are consistent with tags serving as memory cues. If they do serve this purpose, we should expect increased listening rates for musical artists once a tag is applied, under the assumption that a tag facilitates retrieval and increaes the chances of a user listening to a tagged artist. A cursory examination

Total users	104,829
Total scrobbles	2,089,473,214
Unique artists listened	4,444,119
Unique artists tagged	1,049,263
Total user-artist listening time series	94,875,106
Total tagged time series	5,930,594
Total untagged time series	88,944,512

Table 1: Dataset summary

of the data demonstrate that listening rates for tagged time series are much greater than for untagged time series (the average number of listens across all time series is 16.9 when untagged and 98.9 when tagged). While suggestive of the importance of tagging, more involved analysis is required to say whether the difference actually has anything to do with tagging. This could simply indicate that users are much more likely to tag those artists they are likely to listen to more anyway.

Thus we must compare tagged and untagged time series in a "fair" manner, controlling for as many factors as possible. To do so, we first need a means of temporally aligning tagged and untagged time series. Tagged time series are aligned so that they are centered on the month in which they were tagged (allowing us to compare pre- and post-tagging behavior), but there is no direct analog to this among the untagged data. We resolve this by noting that tagging is disproportionately likely in a user's peak listening month for a given artist: In approximately 30% of our 6 million tagged time series, a user has tagged an artist in the same month where she has listened to that artist most. This provides a basis for aligning the tagged and untagged time series, by selecting only those tagged time series where the tag was applied in the month of peak listening, and a matched sample of untagged time series. After aligning both samples to the month with the most listening (i.e. so that all time series are centered about the peak), we limited our analysis to a 13 month period extending from 6 months prior to the peak month to 6 months after the peak. To ensure comparability we constain the time series to have:

- more than 25 total listens;
- a peak in listening at least 6 months from the edges of our data collection perod (i.e. ensuring that the period from 6 months before to 6 months after the peak does not extend beyond the limits of our data range); and
- at least one listen 6 months prior to and after the peak (i.e. if the peak occurs in July 2008, there should be at least one listen in January 2007, and one listening in January 2009).

Constraining our time series in this manner, we were left with a total 206,140 tagged time series, and randomly sampled an equal number of tagged time series meeting the same three criteria.

In Figure 1a we plot the mean playcounts (log-transformed and normalized

Why do we have to log transform this one? Just for comparison, wouldn't the straight (normalized) means be fine?

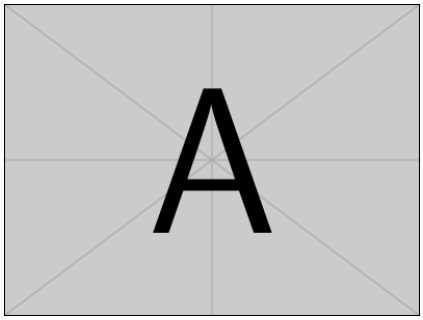
) for tagged and untagged time series for each month in our analysis window. Both show relatively increased listening rates in the months after the peak as opposed to those prior. We observe, however, a small but reliable

Probably need to say something about the CIs here, if we leave them in, or else otherwise justify that this is a meaningful difference

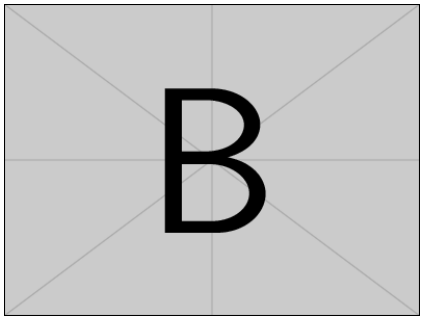
effect wherein tagged time series show proportionally higher listening rates after the peak month (in which the the tag was applied) as compared to untagged time series. This is suggestive of an increase in listening as a result of tagging.

To more robustly test for such an effect, we next developed a generalized additive regression model relating post- and pre-peak listening behavior.

Okay, here's where you can insert your description of the regression mode/results, to make sure I don't muck it up



(a) Mean log-transformed and normalized playcount by month.



(b) Regression results, with 95% confidence interval.

Fig. 1: Comparison of tagged and untagged listening time series

Thus we can conclude that ...

I want a concise, clear nugget of the takeaway from the regression results, but want to be 100% sure on the results first. I'm assuming it will boil down to some thing like "...conclude that tagging leads to a small, but significant effect on future listening amount to ~ 2 more listens in the post-tag period than we would expect in the absence of the tag". Something along those lines

4.2 Tag analysis

To examine if and how different tags are associated with increased future listening, we repeated the same regression analysis (on the tagged time series along) described above, but this time included binary (present / not present) regressors for all unique tags that had at least 10 occurrences in our subsample. Among the $\sim 200,000$ annotations captured by our tagged time series, this amounted to $\sim 1,200$ unique tags. Of these, 108 proved to be statistically significant

What was your p-value here? 0.05? Also not sure how much extra detail on the regression model is needed here, as it's basically the same as what we describe in the previous section

in the resultant mode. While we only have sufficient evidence to make claims about these 108 tags, qualitative examination of which tags are relatively strong predictors in the model proves informative.

The most telling observation is that commonly-used genre tags (e.g. "pop", "jazz", and "hip-hop") – which are the most common tags overall in our full dataset – tend to be weak, negative predictors of future listening. In contrst, relatively strong predictors (both positive and negative) appear to be relatively obscure, possibly idiosyncratic tags ("arguman-loved tracks", "mymusic", "leap-sandbounds cdcollection").

It seems like it would be nice to have a full listing of the tags and coefficients for people to look at, but of course listing it untenable...maybe we link to it in a footnote? Or not worthi it?

To examine this trend quantitatively, we plot in Figure 2 global tag popularity (i.e. the total number of uses of a tag in our full dataset, which consists of ~ 50 million annotations) as a function of its coefficent in our regression model. This reveals a clear trend of the most popular tags being negative, but very weak, predictors of future listening, while strong predictors tend to be relatively unpopular.

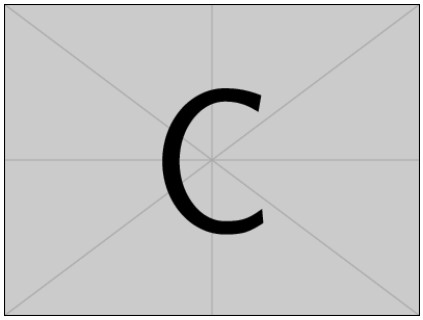


Fig. 2: Tags' global popularity as a function of coefficient in our regression model.

I think we should plot this against global tag popularity (i.e. from the full dataset..probably log popularity though). I doubt the shape will be different, but I think that's more appropriate here. I can do this easiy once I have the final list of significant tags. Agreed? By the time you see this I'll have pushed a file to git with these global frequencies.

These data suggest that, at least for the small number of tags about which we can make statistically meaningful claims, those that are globally popular and well-known have relatively little effect on future listening, and are generally associated with small decreases in post-taggging listening rates. The tags that seem to "matter" (i.e. those that are relatively strong predictors of whether or not a user will listen to an artist after tagging it) are generally much less popular.

5 Conclusion

Need to think through this section a bit more, once results are a bit more finalized, but the following chunk definitely should be in there (though certainly with some edits).

Based on such a small sample, we are at this point tentative to make strong claims about what specifically differntiates those unpopular tags that are strong negative versus strong positive predictors of future listening. The evidence is suggestive of, nevertheless, relatively uncommon (and likely, in many cases, to be idiosyncratic) tags being those most predictive of future listening behavior. This raises the intriguing possibility that the descriptive, popular tags that are arguably most useful to the community at large (i.e. genre labels and related tags), are not particularly strong predictors of future listening, and thus are likely not functioning as memory cues.

This suggest that, while on average tagging an artist has a small positive effect on future listening, the most common tagging activities are *not* are not strong predictors of future retrieval listneing. We cannot be sure which of the many other possible tagging motivations are at play here, nor can we know if and when a tag is applied with the intention of being used for retrieval, while ultimately not being used for this purpose. That said, these results do suggest that descriptive, relatively well-known genre classifiers do not show evidence of use as retrieval aids, but are nonetheless the most commonly applied tags. This may indicate that the primary motivation for tagging on Last.fm is not for personal information management (tagging a resource for one's own retrieval), but rather is socially-oriented, resulting in tags that are useful for the community at larger.

This leads to the interesting possibility that a folksonomy can generate the useful, crowdsourced classification of content that proponents of collaborative tagging extol, but that this process is not strongly driven by the self-directed, retrieval-oritented tagging that is typically assumed in such systems.

Need to talk about limitations, both of methods, and of course that we're talking just about one system...could be different on, say, Delicious, but we don't yet have to test it

Todo list

References

1. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 971–980. ACM (2007)
2. Cattuto, C., Loreto, V., Pietronero, L.: Semiotic dynamics and collaborative tagging 104(5), 1461–1464 (2007)
3. Glushko, R.J., Maglio, P.P., Matlock, T., Barsalou, L.W.: Categorization in the wild 12(4), 129–135 (2008)
4. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems 32(2), 198–208 (2006)
5. Gupta, M., Li, R., Yin, Z., Han, J.: Survey on social tagging techniques 12(1), 58–72 (2010)
6. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. In: Proceedings of the 16th international conference on World Wide Web. pp. 211–220. ACM (2007)
7. Heckner, M., Heilemann, M., Wolff, C.: Personal Information Management vs. Resource Sharing: Towards a Model of Information Behavior in Social Tagging Systems. In: ICWSM (2009)
8. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Proceedings of 3rd European Semantic Web Confernece (ESWC). pp. 411–426. Springer (2006)

9. Körner, C., Benz, D., Hotho, A., Strohmaier, M., Stumme, G.: Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In: Proceedings of the 19th international conference on World wide web. pp. 521–530. ACM (2010)
10. Körner, C., Kern, R., Grahsl, H.P., Strohmaier, M.: Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In: Proceedings of the 21st ACM conference on Hypertext and hypermedia. pp. 157–166. ACM (2010)
11. Lorince, J., Todd, P.M.: Can simple social copying heuristics explain tag popularity in a collaborative tagging system? In: Proceedings of the 5th Annual ACM Web Science Conference. pp. 215–224. ACM (2013)
12. Lorince, J., Zorowitz, S., Murdock, J., Todd, P.M.: "Supertagger" behavior in building folksonomies. In: Proceedings of the 6th Annual ACM Web Science Conference. pp. 129–138. ACM (2014)
13. Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the seventeenth conference on Hypertext and hypermedia. pp. 31–40. ACM (2006)
14. Nov, O., Naaman, M., Ye, C.: What drives content tagging: the case of photos on Flickr. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 1097–1100. ACM (2008)
15. Robu, V., Halpin, H., Shepherd, H.: Emergence of consensus and shared vocabularies in collaborative tagging systems 3(4), 14 (2009)
16. Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J.: Tagging, communities, vocabulary, evolution. In: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work. pp. 181–190. ACM (2006)
17. Vander Wal, T.: Explaining and Showing Broad and Narrow Folksonomies (2005)
18. Vander Wal, T.: Folksonomy Coinage and Definition (2007)
19. Yeung, A., Man, C., Noll, M., Gibbins, N., Meinel, C., Shadbolt, N.: On measuring expertise in collaborative tagging systems (2009)
20. Yeung, C.m.A., Noll, M.G., Gibbins, N., Meinel, C., Shadbolt, N.: SPEAR: Spamming-Resistant Expertise Analysis and Ranking in Collaborative Tagging Systems 27(3), 458–488 (2011)
21. Zollers, A.: Emerging motivations for tagging: Expression, performance, and activism. In: Workshop on Tagging and Metadata for Social Information Organization, held at the 16th International World Wide Web Conference (2007)
22. Zubiaga, A., Körner, C., Strohmaier, M.: Tags vs shelves: from social tagging to social classification. In: Proceedings of the 22nd ACM conference on Hypertext and hypermedia. pp. 93–102. ACM (2011)