Comparing Social Tags to Microblogs

Victoria Lai, Christopher Rajashekar, William Rand
Center for Complexity in Business
University of Maryland
College Park, Maryland, USA
{vlai, wrand}@umd.edu, christopher.rajashekar@rhsmith.umd.edu

***Abstract*—As Internet usage and e-commerce grow, online social media serve as popular outlets for consumers to express sentiments about products. On Amazon, users can tag an album with a keyword, while tweets on Twitter represent a more natural conversation. The differing natures of these media make them difficult to compare. This project collects and analyzes social media data for newly released music albums and develops new methods of comparing a product's social tags to its microblogging data. It explores information retrieval and rank correlation measures as similarity measures, as well as term frequency-inverse document frequency (*tf-idf*) processing. We conclude that with sufficient Twitter activity about an album, social tags do represent the most frequent conversations occurring on Twitter. These results imply that managers can collect and analyze tags and use them as a proxy for most common consumer feedback from microblogging, which is more difficult to collect.**

Keywords-social media; tagging; microblogging; comparison; similarity; music albums; Twitter; Amazon

I. INTRODUCTION

As new albums are released, a convenient way to view consumer feedback is through online media. Sites such as Amazon let consumers rate, review, and tag (generate related keywords for) music albums. Meanwhile, microblogging sites such as Twitter offer a different form of feedback — what are consumers actually saying about the product to their friends?

Though tags serve organizational and information-finding purposes on sites like the LastFM music community and Del.icio.us, a social bookmarking platform, tags on Amazon, a transaction-based site, also provide brand managers with important product review information where purchases are actually taking place. Tags are easier to collect than tweets since there are fewer of them, and Twitter's rate-limiting and terms of service restrictions make tweet collection difficult. So if tags represent Twitter conversations well, managers could use tags as a proxy for Twitter activity.

However, there is no established method to compare the two dissimilar datasets. This project aims to analyze social media data for newly released music albums to (1) develop a framework for discussing methods of comparison between social tags and microblogging data and (2) discuss and compare the results of those methods as applied to the albums.

II. RELEVANT WORK

As opposed to previous work [1, 3, 7, 9] that compared tags to other tags, we propose a similarity framework for comparing tags to tweets where user motivations differ. [8] used structured metadata on the Flickr photo-sharing service to generate better organizational hierarchies, but tweets are unstructured raw data. Previous work in tag prediction and recommendation methods has used information retrieval (IR) [4] and correlation measures [2] to measure similarity between sets of predicted tags and the ground truth. In this paper, though, the datasets are not as similar or clean as two sets of tags; rather, the methods must be applied to a smaller set of tags and a much larger, noisier set of 140-character tweets.

Other previous work [5] compared hashtags, i.e., strings preceded by a hash (#), in tweets to tags on Del.icio.us and found that tagging behavior and motivations differ on the two sites. Hashtags, like tags on Del.icio.us, still involve a conscious decision, so it is still unclear whether the full content of tweets compares to tags in other media. [10] found that aggregating tweets into different streams offers different properties in examining a topic. We aggregate tweets based on message content about a particular album (a keyword stream) but aggregation by some other method such as hashtags or users could provide different results. [6] examines ways to automatically summarize tweets. We show that tags can also serve as a summary of tweets in special cases, but do not require any processing. In order to generate personalized tags for a Twitter user's interests based on her tweets, [11] applied term frequency-inverse document frequency (*tf-idf*) ranking. One of their findings was that the tagging precision of *tf-idf* on tweets is comparable to that of web keyword extraction used for advertising. Their work analyzes tweets grouped by users, while we group tweets by topic and additionally collect background tweets as control data in *tf-idf* analysis.

III. DATASET DESCRIPTION

Twelve albums were selected for analysis, with release dates ranging from 12/7/2010 to 2/22/2011. Using Twitter's Search API, data about each album was queried on a regular basis using album name, artist name and/or words like "album" if needed to filter out irrelevant results. The Search API limits query results to 1500 tweets, so our tool was run every two days to collect as many new tweets over the time period as possible. At the same time, tags and tag weights, their respective number of times tagged, were collected from Amazon on a weekly basis.

Since overall tweet content varies over time, it is important to collect control data to filter out background noise. For this

purpose, we collected data for a music-related control set and a general control set. The *music* control set, consisting of results for the search term "music", provided a baseline of comparison for whether album-specific tags represented album-specific tweets better than they represented generic music-related tweets. For the general control set, we ran searches on common English words such as "I", "the", and "and". *Tf-idf* analysis utilized only the general control set, which was more representative of background noise.

IV. Data Processing

Once the tweets were collected, the most common words and phrases associated with each album were determined. Raw frequency counts were used to sort the tweet keywords but failed to take into account the frequency of each word or phrase on Twitter overall. As a result, *tf-idf* weights were also assigned to the phrases to measure the relative importance of keywords in each document, with searches from the general control set included in the corpus to filter out background noise. The tweets were grouped together by album as individual documents, as were the results from each of the five general control searches, for a total of 17 documents. The *tf-idf* weight of a phrase in a particular document increases with the frequency of the phrase in the document and is inversely related to the total number of documents that contain the phrase. The *tf*, *idf*, and *tf-idf* of a phrase *i* in a document *d* are computed as shown in (1, 2, 3).

$$tf_{i,d} = \frac{n_{i,d}}{\sum_{k=1}^{p} n_{k,d}} \tag{1}$$

$$idf_i = log \frac{|D|}{|\{d : i \in d\}|} \tag{2}$$

$$tf\text{-}idf_{i,d} = tf_{i,d} * idf_i \tag{3}$$

V. Similarity Framework

Social tags consist of user-generated words or phrases, while tweet keywords and phrases are pulled from natural language. In our dataset, the number of Amazon tags and tag weights may be in the hundreds while total tweet phrases and frequency counts can be in the tens of thousands. For each Amazon tag (t_a), we can judge its importance and ranking based on some function $f_a(t_a)$, where the tag weights are the simplest measure. For Twitter, we can extract terms (t_w) and term frequency counts (*freq*), but the question is if there is a better weight to assign each word than just the frequency counts. So we propose different importance measures $f_w(t_w)$, such as term frequency count and *tf-idf* weights (*tf-idf*), to compare to tag weights.

We also propose similarity measures S to evaluate the relationship between $f_a(t_a)$ and $f_w(t_w)$. We check if the result is greater than some threshold θ, which we base on the *music* control dataset. If album-specific tags match album-specific tweets better than they do generic *music* tweets, then we conclude that Amazon tags do provide some insight into Twitter conversations about the product. Our process involves manipulating (1) the tag sets used as t_a and t_w, (2) the importance measures f_a and f_w, and (3) the similarity metric S. We can now express the test for a successful set of importance functions and similarity measures as:

$$S\left(f_a(t_a), f_w(t_w)\right) > \theta. \tag{4}$$

The sets t_a and t_w for tags and tweet keywords can be all that were collected (*all tags*), the top k, the top k with generic product characteristics like artist and album names removed (*queries removed*), or a similar variation. The importance measure f_a for tags is, for simplicity, the tag weight for some t_a. Due to the differing natures and sizes of the tag and tweet keyword sets, we focus on how well Amazon tags reflect Twitter keywords rather than looking at two-way similarity.

A. Correlations (S = C)

Spearman rank correlation compares two ranked datasets to measure the correlation strength between their ranks. Since we focus on tags as a proxy for tweets, we consider the correlations for *all tags* and *top tags* between tag weights and their importance measure as found in tweets (*freq* or *tf-idf*). This addresses the question of whether popular social tags are tweeted more often than unpopular tags. The Kendall tau correlation is another form of rank correlation, which is based on the number of concordant and discordant pairs.

B. Information Retrieval (S = IR)

IR measures such as precision and recall look at how well documents retrieved in a query represent some "ground truth" of relevant documents. While a case could be made for either tags or tweets to be the ground truth, tweets tend to be the more elusive content whereas tags are easily retrieved. As such, an appropriate question is: Are tags representative of what people are saying on the more data-intensive Twitter?

With tags as our retrieved results and tweet keywords as our ground truth, we are able to ask how well Amazon tags serve as an IR mechanism for tweet content about the album. Precision measures how many of the retrieved results are found in the ground truth, so it is the fraction of tags about an album that match tweet keywords. Recall measures how much of the ground truth is found in the retrieved results, or the fraction of tweet keywords matched by tags. There is usually a tradeoff between precision and recall.

VI. Results

We now proceed to determine whether tag content reflects tweet content for the selected music albums. This was done using the proposed variations within the similarity framework, on tag sets, importance measures, and similarity metrics.

A. Correlations

Table I shows the rank correlation coefficients from Spearman and Kendall tau in comparing the tag weights to tweet counts for albums with at least ten tags. The maximum values for each album are highlighted. The threshold is the *music* control set C1 (5), which looks at the tags' weights

TABLE I. SPEARMAN AND KENDALL TAU RANK CORRELATION COEFFICIENTS

Album	C1: t_a = *all tags*, f_w = *freq*, t_w = *music*		C2: t_a = *all tags*, f_w = *freq*		C3: t_a = *top tags*, f_w = *tf-idf*	
	Spearman	*Kendall.*	*Spear.*	*Kend.*	*Spear.*	*Kend.*
D1	0.44	0.38	0.29	0.25	**0.69**	**0.43**
D2	0.29	0.24	0.38	0.37	**0.78**	**0.70**
D3	0.24	0.20	0.38	0.33	0.33	0.31
D4	0.30	0.26	0.40	0.35	**0.60**	**0.51**
J1	0.64	0.55	0.31	0.28	**0.31**	**0.28**
J5	0.20	0.18	0.23	0.18	**0.63**	**0.44**
J6	0.47	0.37	0.28	0.19	**0.63**	**0.45**
F2	0.24	0.20	0.43	0.36	0.30	0.28

against their frequency counts in the *music* control set. We begin with the most intuitive base case C2 (6), where we compare all tag weights to their corresponding frequency counts in album tweets.

$$C1{:}C\big(f(\{\text{all tags}\}), freq(\{\text{music tweets}\})\big) \quad (5)$$

$$C2{:}C\big(f(\{\text{all tags}\}), freq(\{\text{album tweets}\})\big) \quad (6)$$

C2 is stronger than C1 for some albums, meaning with this measure, tags only sometimes represent album tweets better than they represent *music* tweets. Frequency count correlations were also found for *top ten tags* (not shown: t_a = *top tags*, f_w = *freq*) and correlation values were usually higher than those for *all tags*. This supports the conclusion that top-ranked tags are more relevant and less popular tags are less relevant.

Other variations of the tag set and *freq* and *tf-idf* were tried, and it was found that correlations as a similarity metric are most effective when considering the *top ten tags* with *tf-idf* as the importance measure (C3). This combination resulted in the strongest correlations, significantly stronger than the threshold C1, for most albums. *Tf-idf* uses the *tf-idf* weights assigned to phrases against the general control set to determine the relative importance rankings on Twitter. The *tf-idf* weights were meant to filter out background noise, and the *tf-idf* correlations shown in bold are at least as strong as C2 and all except one album are stronger than the threshold C1. So *tf-idf*, in addition to filtering irrelevant results, can improve the correlation between rankings of relevant keywords and rankings of *top tags*.

A variation of C3 is the queries removed set (not shown: t_a = *top tags* \ {*queries*}, f_w = *tf-idf*), in which we exclude tags that were a subset of the search query used to collect the album-related tweets. The correlations are moderately strong for some albums and weak for others, showing that characteristics of the album strengthen the correlation but are not the only factor, except for a few albums. D3 and J1 resulted in negative correlations since low tag activity resulted in only one of the remaining tags to still be found in tweets after query removal. For these two albums, the *music* control set correlations were strongest for *top tags* (not shown: t_a = *top tags*, f_w = *freq*, t_w = *music* tweets), showing that with low tag activity, their tags did not represent the Twitter conversation about the album very well, actually correlating more strongly with *music* tweets.

B. Information retrieval

With IR (7), we look at each album's tags as a retrieval mechanism for each album's tweets. Since tweet keywords in the ground truth are all considered "relevant" and unranked, the importance measure for the tweets is not used.

$$IR\big(f(\{\text{all tags}\}), f(\{\text{album tweets}\})\big) \quad (7)$$

Our threshold is how well each album's tags represent *music* tweets. Shown in Table II are the precision and recall for 10 of the 12 albums, where the remaining two had no Amazon tags. The albums are split into high and low volume (HV, LV) based on number of tweet keywords divided by number of tags. The queries have higher precision when there are more tweet keywords, such as for J5, J6, and F2, and poor precision with low tag activity and/or low Twitter activity, such as for D2, J3, and F1. The HV average for the precision values (P1) is greater than for threshold precision values (P2) from *music*, while the LV average is greater for P2. Thus tags reflect album tweet content more accurately when there is more Twitter activity.

Individual album comparison of P1 to P2 reveals interesting results. Albums with P1 less than D4's P1 of 0.36 end up with higher P2, albums with P1 greater than 0.36 end up with lower P2, and D4's P1 and P2 values are equal at 0.36. This seems to suggest that for albums with low tag or Twitter activity, the tags reflect *music* tweets better than they reflect specific album tweets, and for albums with sufficient Twitter activity, the album tags reflect specific album tweets better. These findings are consistent with those from the correlation analysis, where the *top tags* for albums with low tag activity correlated better with generic *music* tweets than with

TABLE II. INFORMATION RETRIEVAL MEASURES WITH AMAZON TAGS AS QUERIES AND ALL TWEET KEYWORDS AS GROUND TRUTH

Album	Precision (P1)	Precision threshold (P2)	Recall
D1	0.48	0.43	0.002
D2	0.24	0.62	0.008
D3	0.29	0.36	0.001
D4	0.36	0.36	0.0004
J1	0.20	0.50	0.0003
J3	0.00	0.75	0.00
J5	0.57	0.40	0.0002
J6	0.75	0.38	0.0004
F1	0.00	0.50	0.00
F2	0.67	0.59	0.00009
Average	**0.35**	**0.49**	**0.001**
HV average	**0.51**	**0.45**	**0.0003**
LV average	**0.20**	**0.53**	**0.002**

album-specific tweets.

Recall values are low because the tags are small in number while the set of tweet keywords is much larger. This raises the possibility of a representation bias in the tags. To test this, we compared the *idf* distribution of all single-word album tweet keywords (single-word to remove the skew caused by phrase variations) to that of all tags that matched album tweets (*tag-match*) using the Kullback-Leibler (KL) divergence measure, which calculates the difference between two distributions. The KL from *tag-match* to the album tweet keywords, the ground truth, was 0.20. The KL from *tag-match* to the general and *music* controls was 0.41 and 0.39, respectively. The much smaller KL of *tag-match* to the album keywords means that the distributions are relatively similar, suggesting that the tags that match are an unbiased sample of album tweet keywords.

Finally, to determine the effects of changing the tag set to *top tags* instead of *all tags*, we can look at precision-recall curves. Precision-recall curves plot the precision and recall for the top-*k* as *k* goes from 1 to the total number retrieved. As shown in Fig. 1, precision is high at the beginning of the precision-recall curves and then drops sharply for most albums. We can see that, as shown in the correlations as well, top-ranked tags are relevant to the Twitter conversations, whereas lower-ranked tags tend to be less relevant.

VII. Conclusions and Future Work

Given our findings from the IR and correlation measures, it does appear that Amazon tags serve as a good proxy for top Twitter conversations when there is sufficient consumer interest in tagging and generating Twitter content. The tags not only match tweet keywords with high precision in those cases but are also representative of the entire set of tweet keywords. If only interested in the most important Twitter trends for new albums, brand managers can primarily consider social tag content. Our method does not work very well for niche artists where the audience might not be large enough to generate sufficient tagging or Twitter activity.

Especially when there is more Twitter activity, low recall shows the tags do not capture everything. The more relevant tags are near the top in tweet keyword rankings, as shown by the precision-recall curves and rank correlations. So *top tags* can serve as a proxy for top keywords (album title, artist, genre, etc.), but will not capture entire Twitter conversations about the album. *Tf-idf* processing does help to filter out background noise and can improve correlation results.

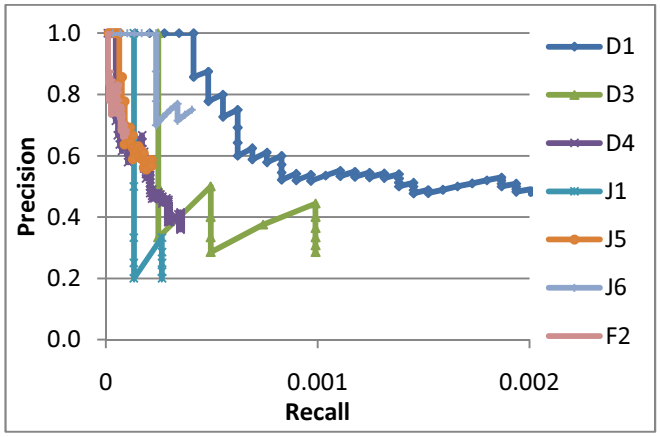

Figure 1. Precision-recall curves for seven albums, with D2 excluded for scaling reasons

A larger dataset with more products and tag activity is necessary to further confirm these hypotheses, and removing the requirement of an album's recent release could realize this. Nonetheless, data on new albums allow for comparison of tags and tweet keywords over time after product release, and indicate that this technique works even for new products. The analyses could also be applied to tags from other sources, specifically LastFM. Categorizing tags and tweet keywords by characteristics such as purpose and sentiment may reveal interesting results as to which online medium consumers tend to use for particular kinds of feedback. Finally, linguistic analysis such as clustering and stemming would help in filtering and grouping tweet keywords.

Acknowledgment

We thank the National Science Foundation (Award #1018361) and the Center for Complexity in Business for supporting this research.

References

[1] M. Clements et al., "Detecting synonyms in social tagging systems to improve content retrieval," in *Proc. of the 31st Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Singapore, 2008, pp. 739-740. doi: 10.1145/1390334.1390479

[2] D. Eck et al. (2007). *Automatic generation of social tags for music recommendation* [Online]. Available: http://www.iro.umontreal.ca/~eckdoug/papers/2007_nips.pdf

[3] E. Giannakidou et al., "SEMSOC: SEMantic, SOcial and Content-based clustering in multimedia collaborative systems," in *Proc. of the 2008 IEEE Int. Conf. on Semantic Computing*, Santa Clara, CA, 2008, pp. 128-135. doi: 10.1109/ICSC.2008.73

[4] P. Heymann et al., "Social tag prediction," in *Proc. of the 31st Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Singapore, 2008, pp. 531-538. doi: 10.1145/1390334.1390425

[5] J. Huang et al., "Conversational tagging in Twitter," in *Proc. of the 21st ACM Conf. on Hypertext and Hypermedia*, Toronto, 2010, pp. 173-178. doi: 10.1145/1810617.1810647

[6] D. Inouye and J. K. Kalita, "Comparing Twitter Summarization Algorithms," in *IEEE SocialCom 2011*, Boston, MA, 2011.

[7] B. Markines et al., "Evaluating similarity measures for emergent semantics of social tagging," in *Proc. of the 18th Int. Conf. on World Wide Web*, Madrid, 2009, pp. 641-650. doi: 10.1145/1526709.1526796

[8] A. Plangprasopchok et al., "Growing a tree in the forest: constructing folksonomies by integrating structured metadata," in *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Washington, DC, 2010, pp. 949-958. doi: 10.1145/1835804.1835924

[9] R. Schenkel et al., "Efficient top-k querying over social-tagging networks," in *Proc. of the 31st Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Singapore, 2008, pp. 523-530. doi: 10.1145/1390334.1390424

[10] C. Wagner and M. Strohmaier, "The wisdom in tweetonomies: acquiring latent conceptual structures from social awareness streams," in *Proc. of the 3rd Int. Semantic Search Workshop*, Raleigh, NC, 2010. doi: 10.1145/1863879.1863885

[11] W. Wu et al., "Automatic generation of personalized annotation tags for Twitter users," in *Human Language Technologies: The 2010 Annu. Conf. of the North American Chapter of the Assocation for Computational Linguistics*, Los Angeles, CA, 2010, pp. 689-692.