#nowplaying the Future Billboard: Mining Music Listening Behaviors of Twitter Users for Hit Song Prediction

Yekyung Kim
Music and Audio Research Lab
Seoul National University
Seoul, South Korea
peiru@snu.ac.kr

Bongwon Suh
Human Centered Computing Lab
Seoul National University
Seoul, South Korea
bongwon@snu.ac.kr

Kyogu Lee
Music and Audio Research Lab
Seoul National University
Seoul, South Korea
kglee@snu.ac.kr

ABSTRACT

Microblogs are rich sources of information because they provide platforms for users to share their thoughts, news, information, activities, and so on. Twitter is one of the most popular microblogs. Twitter users often use hashtags to mark specific topics and to link them with related tweets. In this study, we investigate the relationship between the music listening behaviors of Twitter users and a popular music ranking service by comparing information extracted from tweets with music-related hashtags and the Billboard chart. We collect users' music listening behavior from Twitter using music-related hashtags (e.g., #nowplaying). We then build a predictive model to forecast the Billboard rankings and hit music. The results show that the numbers of daily tweets about a specific song and artist can be effectively used to predict Billboard rankings and hits. This research suggests that users' music listening behavior on Twitter is highly correlated with general music trends and could play an important role in understanding consumers' music consumption patterns. In addition, we believe that Twitter users' music listening behavior can be applied in the field of Music Information Retrieval (MIR).

Categories and Subject Descriptors

H.2.8 [Database Application]: Data mining

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Social media mining, music listening behavior, hit song science, Random Forest, Support Vector Regression, machine learning

1. INTRODUCTION

As the use of microblogs substantially increases, abundant data becomes available online. The information shared in microblogs enables researchers to investigate the various activities and thought of users.

Music-related activities are no exception. People share rich information about their musical activities (e.g., listening to a particular music or discussing songs and artists). While such information is available on social media, there has been limited research on how to take advantage of the musical activities available on social media for the benefit of MIR (Music Information Retrieval) [5]. Online music services such as Last.fm provide this data with their public API[1], but its use is limited while there could be bigger opportunities if this data is available publicly. The bulk of user related music information on such online music services is difficult to access for general purposes.

Music consumption and distribution platforms change from offline to online with the growth of the online market size and the digitalization of music content. The growth of digital music sales increases continuously and overtook physical music sales in 2012 according to a Nielsen report[2]. It implies that users' online activities such as music listening behavior become important data for the music industry. We believe that users' behavior related to music on online platforms is significantly associated with music sales. The information about when and what sort of music a user has listened to is valuable for understand the patterns of user music consumption and can be applied in various fields.

In this paper, to model the users' music listening behavior, we collected tweets from Twitter, which contain a set of music-related keywords (e.g., #nowplaying and #itunes). In Twitter, even though it is not regulated, it is an established practice that such hashtags are used to share what music a Twitter user is playing at that moment. As a matter of fact, #nowplaying is widely used on Twitter, as shown in Suh et al. [12] where a number of #nowplaying hashtags were in the top of the hashtag rank. However, it is uncertain that music listening behaviors of Twitter have meaningful relationships with general music trends since Twitter users do not necessarily represent the general public.

In this study, we investigate the relationship between music related activities in Twitter and popularity in the general music market. This approach explores the possibility that the Twitter dataset could be used as a reflection of general music trends. In addition, using the features extracted from music listening behavior of Twitter, we suggest a method that predicts whether the music will be a hit or not on the next week's Billboard chart. Our approach uses a number of music related features in tweets including music, artists, and the number of weeks a song has been listed on the Billboard chart. In this paper, we evaluate the applicability of music listening behavior acquired in tweets by

---

[1] http://www.last.fm/api

[2] http://www.nielsen.com/us/en.html

how well our features could predict future Billboard ranks.. The research could play an important role in understanding consumers' music consumption patterns. In addition, predicting whether songs are a hit or not on the Billboard is a matter of concern not only for the music market but also in MIR. We believe that our approach to predict a hit song using music listening behavior could inform the improvement of the performance of hit song prediction.

2. RELATED WORK

Our research can be categorized into the retrieval of music-related information from microblogs and hit prediction using the social media data.

Recently, the "Million Musical Tweet Dataset" (MMTD), a dataset of music listening activities collected from Twitter with music-related hashtags was released [4]. It is the biggest publicly available dataset extracted from microblogs that includes geographic, contextual information, and genres. Goel et al. [3] crawled a 1% sample of total tweets and only retained 3% of the collected tweets that had geographic information to focus on geospatial information.

In this paper, our dataset consists of data crawled with a Twitter Streaming search API that returns real-time public statuses. Specifically, it returns all the tweets that contain a certain keyword. Using this, instead of collecting only 1% sample of tweets, we were able to collect all tweets containing a set of music-related keywords.

With their dataset of music listening behavior, Schedl [8] analyzed geospatial music listening patterns and estimated artist similarities and their popularity. Using the geospatial notions of similarity, Schedl and Schnitzer [10] proposed hybrid music recommendation systems and showed that adding a geo-location feature benefits the performance of collaborative filtering.

For hit prediction using social media, researchers focused on the use of metadata available in the social media data. Pachet and Roy [7] claim that the popularity of a song cannot be only learnt by its intrinsic characteristics. In Bischoff et al. [1], the proposed method predicts the potential of tracks to becomes the hit songs. It uses the social data that are mined from the music social network site 'Last.fm' investigating the relationship between tracks and metadata. The work evaluates the performance of prediction by dividing the same size of hit group and non-hit group. Meanwhile, the use of P2P search queries was proposed by Koenigstein et al. [6]. They found a strong correlation between P2P queries and Billboard charts. Goel et al. [3] find that the Yahoo search query shows the good performance in predicting the actual rank of Billboard. Furthermore, a combined model that includes both search query and previous Billboard rank performs significantly better. On the other hand, in this research, we utilize a publicly available tweets and the number of weeks a song has been listed on the Billboard chart to quantify the performance of hit prediction. To the best of our knowledge, our work is the first to predict song hits on Billboard charts with music-related hashtags in Twitter.

3. DATASET AND PREPROCESSING

We collected two datasets for the analysis in this study.

- **Music listening behavior dataset:** a dataset of tweets collected with #nowplaying, #np, and #itunes keywords via

Twitter Streaming API [3] (Application Programming Interface) filter over 10 weeks.

- **The Billboard Hot 100**[4]: the most popular single songs across all genres in one week. Radio airplay, sales data, and streaming activities form the basis of the chart ranking.

3.1 Music Listening Behavior Dataset

To crawl a dataset potentially related to music listening behavior, we used a Twitter Streaming search API from November 21, 2013 to January 29, 2014, yielding 31,605,234 tweets. Using the three keywords (#nowplaying, #np, and #itunes), we collected tweets that contain any of the given keywords from public statuses. To verify the collected dataset, we compared the size of our dataset with the statistics provided by "Topsy[5]" a real-time search engine service that analyzes every tweet using the Twitter Firehose API. The results showed that there is only a slight difference about 10% between the data. Thus, we claim that our dataset contains almost all data that contained music-related hashtags, #nowplaying, #np, and #itunes.

#nowplaying is used to model what music users are playing at that moment since it is an established practice among Twitter users. Also '#itunes' is utilized to cover iTunes music listener [9], representing the listening behavior of users with iTunes. Using the #nowplaying and #np keyword, 30,663,504 tweets (97% of whole tweets) were collected and with the #itunes keyword 941,730 tweets (3%) were collected. It should be noted that our dataset is considerably larger than that used in Hauger and Schedl. [5]. However, in spite of the size of the dataset, there are music-related tweets that are not associated with the keywords used in this study. Therefore, we recognize that our dataset is only a part of the total user music listening behavior available in Twitter. Identifying various keywords associated with music listening behavior in Twitter is an important part of future work.

3.2 The Billboard Hot 100

To evaluate the prediction performance, we collected the Billboard Hot 100 chart by Billboard magazine for ten weeks. The Billboard Hot 100 chart is issued every Thursday and lists the 100 most popular current songs in the United States. They describe the ranking of the chart as follows:

" The week's most popular current songs across all genres, ranked by radio airplay audience impressions as measured by Nielsen BDS, sales data as compiled by Nielsen SoundScan and streaming activity data from online music sources tracked by Nielsen BDS. Songs are defined as current if they are newly-released titles, or songs receiving widespread airplay and/or sales activity for the first time." [4]

This data is a good ground truth for representing general trends in the music market and the popularity of songs. The data were acquired from Billboard Biz (http://www.billboard.com/biz) and include artist, rank, title, and the number of weeks on the chart.

Overall, the information about 178 unique songs and 134 artists were collected from the Billboard chart during the same period as the Twitter data.

---

[3] http://developer.twitter.com

[4] http://www.billboard.com/charts/hot-100

3.3 Preprocessing

After collecting tweets with music related keywords, we calculate the song popularity as song play-count, the number of tweets associated with a song. We also count the artist popularity as the number of tweets associated with an artist. To identify a tweet associated with a specific song or artist, we use a full-text search with the song title and the name of the artist. However, succinct and common titles (e.g., "Radio" by Darius Rucker) caused the search result to contain irrelevant tweets. In addition, for songs with the same song title but sung by different artists, when a tweet contained only song titles, the artist is harder to discern. Thus, we excluded 10 songs from the experiment for these reasons. The exclusion resulted in removing 56 entries from 1000 rank entries (songs listed in the Billboard Hot 100 chart during the ten weeks). We used the total of 944 song-rank entries for the analysis. We collected 1,806,438 song-relevant tweets and 3,309,811 artist-relevant tweets for 168 songs with entries in the 10-Week Billboard chart and 124 artists. Since we adapted a full-text search in MyISAM[5] engine with a Boolean mode, tweets with typos are ignored (e.g., Katy perry for Katy parry).

Using the above search method, we compiled the daily play-count of each song and the popularity of each artist. Since the Billboard chart is announced weekly on Thursday, we would have seven daily play-counts and artist popularity scores spanning from Thursday of the previous week to Wednesday of the current week. We chose the median value for each feature to predict the Billboard rank of the current week. Additionally, we added the period of time the song stayed on the Billboard chart after its first entry.

Using this information, we investigated the extent to which song and artist popularity in tweets could be used to effectively predict the Billboard rank in the future.

4. ANALYSIS

Our research goal was to observe whether the information extracted from music listening behavior on Twitter would have a relationship with general trends in the music market and could reasonably predict a song's future ranking. We used three methods of analysis to obtain a quantitative evaluation: (1) correlation measurement, (2) regression model, and (3) classification.

In this section, we examine the correlation between music-related popularity on Twitter and Billboard ranking, and evaluate the predictive model that regress the future Billboard rank with various feature and classify the hit song and non-hit song by hit interval.

4.1 Correlation Measurement

In order to measure the correlation between play-count and Billboard ranks, we calculated Pearson's correlation coefficients between song's play-count, artist popularity and the number of weeks on Billboard chart. The Pearson correlation is used to show a linear relationship between two datasets. The correlation coefficient ranges between -1 and 1, where l, 0, and -1 indicate positive correlation, no relation, and negative correlation, respectively. The convention that a lower rank number is higher rank in the chart (e.g. rank #1 is the highest rank) often makes it confusing when it is compared with play-count where a bigger

---

[5] http://dev.mysql.com

Table 1: Correlation coefficient of features with rank

Feature	Correlation Coefficient
Log(Song Play-count)	0.5592
Log(Artist Popularity)	0.3435
Weeks on Billboard Chart	0.4104

number is indicative of high popularity. To avoid these ambiguities, the rank scale is inverted. For instance, the rank #1 converts to rank score 100, and the rank #100 converts to rank score 1. The linear transformation of rank score for billboard rank is the following:

- $rank_{score} = 101 - rank_{billboard}$

It does not effect on correlation but flips the sign of value. Table 1 shows the correlation coefficients between the rank score and our features. The song play-count has a stronger correlation and the artist popularity has a modest correlative value. In addition, the number of weeks on the Billboard chart feature has a moderately positive association with future ranks. These results indicate that all the attributes we selected correlate with the Billboard rank that represents music trends. In particular, the song play-counts extracted from the music listening behavior dataset show meaningful correlation among the attributes.

4.2 Music Rank Prediction

With the three features collected from tweets, we built three regression models and evaluated the performance of prediction: linear regression, quadratic linear regression, and support vector regression (SVR)[11]. The quadratic linear regression model contains the linear terms, interactions and squared terms while linear regression contains the linear terms only. SVR model uses the radial basis function (RBF) for kernel. Model fitness was measured by a squared correlation coefficient (r^2) between predicted and actual ranks. The set of data was split into training and test sets for SVR using 5-fold cross validation. We evaluated the performance of the regression models.

First, we investigated the performance of the number of weeks on chart feature in predicting song ranks. As shown in Table 2, the SVR model using the feature shows a marginal prediction performance (r^2=0.29). The model predicted the majority of the songs to be in the range of 40–50. Then, we evaluated a Twitter predictive model where the log values of the song play-count and artist popularity were used as features for prediction. The SVR model achieves a considerably high performance (r^2=0.57).

Finally, we built a predictive model with all features: the number of weeks on chart and features extracted from Twitter. The combined model improved the performance from 0.57 to 0.75. This result implies that the number of weeks on chart feature complements the Twitter features. As shown in Figure 1(c), the combined model shows a significant improvement in accuracies.

In summary, we observe that the features extracted from music listening behavior on Twitter could be effectively used in future ranking prediction, whereas the number of weeks on chart feature is marginally correlated. The combined model with the number of weeks on chart data improves the prediction performance significantly.

**Table 2: Squared correlation coefficient for each model that predicts the next week's Billboard rank.**

Squared Correlation Coefficient (R^2)			
	Model		
	Linear regression	Linear regression – quadratic	Support Vector Machine
# of Weeks only	0.17	0.26	0.29
Twitter only	0.32	0.40	0.57
Twitter + # of Weeks	0.40	0.54	0.75
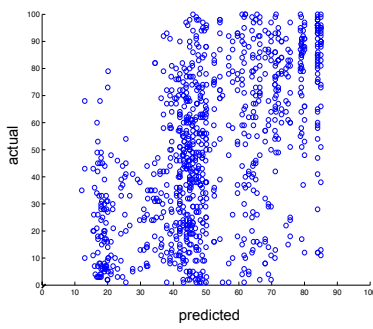
(a) Model with the number of weeks on chart.
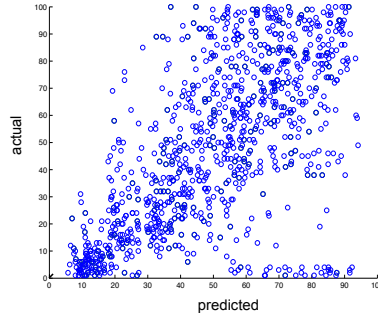
(b) Model with Twitter data that includes song play-count and artist popularity.
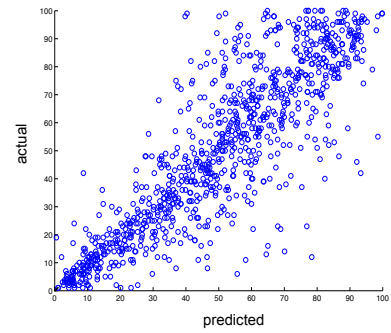
(c) Model combining Twitter and the number of weeks on chart.

Figure 1: Predicted values and actual Billboard rank with SVR model.

With respect to the regression, both linear regression models show quite low performance, indicating that the distribution of features does not sufficiently fit the linear model.

4.3 Music Hit Prediction Algorithm

Our hit prediction classifier is trained with features extracted from Twitter music listening behavior and the number of weeks on chart information. We experimented with a random forest classifier [2] that constructed a multitude of decisions and evaluated its performance by accuracy, precision, recall, and F1-measure for hits/non-hits. It also used a 5-fold cross validation.

The Billboard rank has 100 ranks that remain unchanged for a week, and it is important to define what is a hit versus a non-hit song for the experiment. In this paper, we used several different criteria for hit songs: top 10, 20, 30, 40, and 50. Similarly, we randomly selected the same number of non-hit songs from the

Billboard chart dataset that excluded hit songs. For instance, when the rank 1–10 is the interval of hit, non-hit songs are randomly selected from the 11–100 ranks. Since the dataset used in this study spans ten weeks, we have the total of 1000 song-rank entries. For each week, we chose songs in rank 1-10 as hits and selected randomly chosen from rank 11-100 to non-hits. It resulted in 100 records of hit song (rank 1-10 during the ten weeks) and 100 records of non-hit song (randomly chosen from rank 11-100 during the ten week) as shown in the first row of Table 3. Similarly, we prepared datasets for each hit song criteria as shown in Table 3. Since some songs are excluded from the study due to ambiguity (as explained in the section 3.3), a slightly lesser number of hit songs are used for hit song criteria 1-30, 1-40, and 1-50.

We divided the class of hits and non-hits to have the nearly same size, as in [1]. Classifiers are trained on both sets of hit and non-

Table 3: Evaluation of random forest classifier to predict Hits and Non-hits by various intervals for hit prediction

Hit's Range	# Hit Song	# Non-Hit Song	Accuracy	Hit			Non-Hits		
				Precision	Recall	F1-measure	Precision	Recall	F1-measure
1–10	100	100	0.9	0.92	0.885	0.901	0.88	0.922	0.899
1–20	200	200	0.882	0.91	0.863	0.885	0.855	0.905	0.879
1–30	292	292	0.881	0.88	0.886	0.882	0.883	0.881	0.881
1–40	385	373	0.881	0.86	0.901	0.88	0.903	0.866	0.885
1–50	483	461	0.839	0.83	0.853	0.841	0.848	0.828	0.837

hit songs, according to each interval of the hit range. The features of the songs consisted of the features that showed the best performance in the regression model, song play-count, artist popularity, and the number of weeks on chart.

Table 3 shows that using the hit range 1–10 achieved the highest accuracy. Hit number, non-hit number indicates the number of hit songs and non-hit songs used for classification respectively. It achieved 90% accuracy, 92% precision, 88.5% recall, and a 0.902 F1-measure for hits.

As the range of hit song increase, the gap between hit songs and non-hit songs decreases, making it difficult for the classifier to distinguish hit songs. As shown in Table 3, the accuracy of prediction decreases as wider hit song criteria is employed. With the 1–50 hit range, we classified 944 song-rank entries and obtained 83.9% accuracy, 83% precision, 85.3% recall, and 0.841 F1-measure.

5. CONCLUSION

In this paper, we investigated the relationship between music listening behavior in Twitter and the Billboard rankings. We found that the play-counts extracted from tweets have strong relationships with the Billboard rank, whereas, interestingly, the artist popularity extracted from tweets has a weak correlation with future chart rankings. In addition, the number of weeks on chart information alone was insufficient to predict rank alone.

With the features extracted from tweets, we built three regression models to predict the ranking. Among the proposed models, SVR shows the highest squared correlation coefficient (0.75). Although the combined model with the number of weeks on chart performed the best in rank prediction, the music listening behavior available in Twitter can generate an outstanding predictive model.

We also build a hit prediction classifier with the features acquired in tweets and the number of weeks on chart. We classified the hit and non-hit songs in the Billboard Hot 100 and obtained a value of 83.9% accuracy, 83% precision, and 85.3% recall for classifying a hit song over the whole dataset. The proposed feature showed a high performance both for rank prediction and hit classification. The previous week's twitter features and the number of weeks on chart are effective for predicting the Billboard rank of a song.

In this paper, we were motivated to utilize the music listening behavior from microblogs because it is only a few information available, and the data can be used in various fields. We also investigate the use of the information to improve hit song prediction, which has received a great attention in the music industry and the field of MIR. Analyzing music listening behaviors is crucial to discovering consumer's music consumption patterns and hidden loved songs. We believe tweets could be a meaningful asset to a user-based music recommendation system.

As future work, we plan to investigate additional features regarding music listening behavior, which can be extracted from tweets. In addition, we are interested in analyzing the music listening pattern of individual Twitter users. Profiling Twitter users based on their musical preference and activities would provide interesting insights.

REFERENCES

[1] Bischoff, K., Firan, C. S., Georgescu, M., Nejdl, W., & Paiu, R. (2009). Social knowledge-driven music hit prediction. In *Advanced Data Mining and Applications* (pp. 43-54). Springer Berlin Heidelberg.

[2] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.ISO 690

[3] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). What can search predict. *WWW'10*.

[4] Hauger, D., Kepler, J., Schedl, M., Košir, A., & Tkalcic, M. (2013, November). The Million Musical Tweets Dataset: What Can We Learn From Microblogs. In*Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR), Curitiba, Brazil*.

[5] Hauger, D., & Schedl, M. (2012, October). Exploring Geospatial Music Listening Patterns in Microblog Data. In *Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval (AMR), Copenhagen, Denmark*.

[6] Koenigstein, N., Shavitt, Y., & Zilberman, N. (2009, December). Predicting billboard success using data-mining in p2p networks. In *Multimedia, 2009. ISM'09. 11th IEEE International Symposium on* (pp. 465-470). IEEE.

[7] Pachet, F., & Roy, P. (2008). Hit Song Science Is Not Yet a Science. In *ISMIR*(pp. 355-360).

[8] Schedl, M. 2013 Leveraging microblogs for spatiotemporal music information retrieval. *Advances in Information Retrieval*. 796–799. Springer Berlin Heidelberg.

[9] Schedl, M., & Hauger, D. (2012, April). Mining microblogs to infer music artist similarity and cultural listening patterns. In *Proceedings of the 21st international conference companion on World Wide Web* (pp. 877-886). ACM.

[10] Schedl, M. and Schnitzer, D. 2014 Location-Aware Music Artist Recommendation. In *MultiMedia Modeling* 205–213. Springer International Publishing.

[11] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression.Statistics and computing, 14(3), 199-222.

[12] Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010, August). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*(pp. 177-184). IEEE.