Predicting Billboard Success Using Data-Mining in P2P Networks

Noam Koenigstein Yuval Shavitt Noa Zilberman
School of Electrical Engineering
Tel-Aviv University, Israel
Email: {noamk, shavitt, noa}@eng.tau.ac.il

***Abstract*—Peer to Peer networks are the leading cause for music piracy but also used for music sampling prior to purchase. In this paper we investigate the relations between music file sharing and sales (both physical and digital) using large Peer-to-Peer query database information. We compare file sharing information on songs to their popularity on the Billboard Hot 100 and the Billboard Digital Songs charts, and show that popularity trends of songs on the Billboard have very strong correlation (0.88-0.89) to their popularity on a Peer-to-Peer network. We then show how this correlation can be utilized by common data mining algorithms to predict a song's success in the Billboard in advance, using Peer-to-Peer information.**

I. Introduction

Peer-to-peer (P2P) file sharing is one the most popular activities on the Internet. Despite the high profile legal cases against users and vendors, the exponential growth of users and traffic remains indissoluble. Some ISPs report that file sharing produce more traffic then any other application in the Internet, and copyright owners are advised to start developing business models that will allow them to generate revenue from P2P activity.

Bhattacharjee *et al.* [1], [2] pioneered such directions , where it was suggested that P2P activity can be used to predict an album's life cycle and trends on the Billboard's top 200 albums chart. Both papers used the WinMx file sharing network. In [1] they showed that P2P sharing activity levels provide leading indicators of the direction of movement of albums on the Billboard charts, while in [2] a linear regression model was used to show that sharing activity may be used to predict an album's life cycle. Both of these papers describe a proof of concept, rather than an actual technique or an algorithm.

In [8] we took a different approach and used P2P queries to detect unfamiliar emerging artists; a work that received a great deal of interest in the popular media. In [7] we started investigating the relations between the Billboard and file sharing, where we focused only on the Hot 100, and suggested a novel approach for songs ranking based on piracy. In this paper we take these previous studies a few steps further. Using information we gathered from the popular Gnutella network, we show how well known data-mining algorithms can make use of file sharing information for intelligent decision making that will benefit the music industry. The amount of queries collected (185,598,176), makes it one of the largest P2P mining efforts ever performed (Section II). Using cross-correlation and ranking analysis, we investigate the relations between P2P sharing and actual sales (Section III). We then suggest means to predict an album's top rank on the Billboard that is based on common data-mining algorithms (Section IV).

II. Data-Sets and Methodology

We use three data sources for this study:

- **P2P Search Queries**: A data-set of queries collected from the Gnutella file-sharing network over twenty three weeks.
- **The Billboard Hot 100 Chart**: Hot 100 is the United States music industry standard singles popularity chart. Chart rankings are based on airplay and sales and published weekly by the Billboard Magazine.
- **The Billboard Digital Songs Chart**: Top-downloaded songs across all genres, ranked by sales by the Billboard Magazine.

A. P2P Search Queries

Queries in a file sharing network represent their users current taste and interests. A query is issued upon a request by a user searching for a specific file, or content relevant to the search string. In this study we used data collected from the Gnutella network using the Skyrider systems[1]. According to [10], Gnuella was the most popular file sharing network in the Internet at the time of data collection with a market share of more than 40%. Gnutella is also among the most studied P2P networks in the literature [5], [6], [12]. It is mainly used for piracy of music. In [8] the top 500 most popular queries were manually classified, and it was found that 68% of the queries were music related. Together with adult content (22%), these two categories dominate the query traffic, accounting together for 90% of the queries. Our data set collection period spanned from January 7th 2007 to July 27th 2007 (30 weeks). The total number of US originated query strings processed in this study is **185,598,176**. Our data-set and the technical details of the methodology used to collect it are described in more depth in [5] and [8]. Our data-set is much larger in volume and time span than the one used in [1], [2]. Furthermore,

[1] Skyrider was a startup company that developed file sharing applications and services. The data-set was made available for academic research before the company was closed down. To get access to the data set used in this paper, please contact the authors.

978-0-7695-3890-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ISM.2009.73

IEEE computer society

Rank	**String**	**Occurrences**
1	adult	41,941
2	akon	26,951
3	lil wayne	13,957
5	this is why i'm hot	11,919
6	justin timberlake	11,819
4	beyonce	11,188
7	porn	10,393
8	don't matter	10,259
9	fergie	10,177
10	fall out boy	9,590

Table I
P2P POPULARITY CHART FOR WEEK 7 OF 2007

while Bhattacharjee *et al.* [1] followed the number of channels provided by the sharer and the length of queue waiting to download, we look at users' queries. It seems that [1] did not take into account the possibility of a single user waiting on duplicate sharers' queues; a problem that does not exits when search queries are considered.

B. *The Billboard Charts*

The Billboard charts are the United States music industry standard popularity ranking issued weekly by Billboard magazine. The rankings is based on radio plays and sales (both physical and digital) data collected 10 days before the chart is released. The ranking process does not take into account file sharing activity. New charts are compiled and officially released to the public each Thursday. The charts are dated with the week number of the Saturday after, but in this study we used dates and week numbers according to the actual release date of the chart, and ignored the date issued by Billboard magazine. To simplify time tracking in this paper, we use week numbers instead of full date to chronologically order the Billboard charts and the weekly file sharing data we collected. For example, the Billboard chart which was released on Thursday January 11th 2007 (week number 2), was dated by billboard to January 20th (week 3) but by us to week number 2. A statistical model of songs ranking in the Hot 100 chart can be found in [3].

III. RANKING TRENDS RELATIONS

The Billboard charts rank songs relative to each other, and do not reveal the actual number of sales or air-plays measured during that week. In order to compare it to our file-sharing data, we compiled our own weekly P2P popularity charts based on the popularity of search strings. We measured the popularity of each string by aggregating the number of appearances intercepted from a US based origin during that week.

Table I shows the top 10 positions of the P2P chart generated on week 7 of 2007 (sampled on February 18, 2007). Obviously, the P2P charts include many non music related strings. The string "adult" for example, was ranked number one on every chart we compiled. Unlike the Billboard charts, the P2P charts included also artists names (e.g., Akon or Justin Timberlake), and sometimes even different variations of the same string. In order to avoid inaccuracies, we matched the Billboard songs titles' with their exact match on the P2P chart. Because of the unrelated strings on P2P charts, songs titles always have a lower position on the P2P chart. For example, the song *Before He Cheats* by Carrie Underwood was ranked 17 on the Billboard Hot 100 in the second week of 2007, and 374 on the P2P chart of that week. We thus compiled large P2P charts of at least 2000 strings.

A. *Correlation Measurements*

We want to measure the correlation of trends between different popularity charts. We define $\overline{A_s}$ and $\overline{B_s}$ to be the chart vectors representing the song s on the popularity charts A and B, respectively.

$$\overline{A_s} = \{a_s(1), a_s(2), ..., a_s(n)\} \tag{1}$$

$$\overline{B_s} = \{b_s(1), b_s(2), ..., b_s(n)\} \tag{2}$$

Where $a_s(w)$ and $b_s(w)$ are the positions of song s on charts A and B in week w, respectively. If song s was not in the a chart, we set its position to ∞ for that week. The *support* of a chart vector is the time range that the song was ranked in the chart. Namely for chart A, the support is the set of weeks where $a_s(w) < \infty$. The *joint support* of a song s on charts A and B is the time range in which it simultaneously ranked in both charts.

When a song exits the Billboard charts, it does not mean it is not being played on the radio or sold in stores. Similarly, when a song exits the P2P chart, it does not mean it is no longer being downloaded. Therefore, when considering the correlation of trends between the two charts, we used only the joint support of both charts. Hence we slightly altered the standard definition of cross-correlation to consider only the joint support:

$$corr = \frac{\sum_{i=w_s}^{w_e} [(a_s(i) - E\{\overline{A_s}\}) \cdot (b_s(i) - E\{\overline{B_s}\})]}{\sqrt{\sum_{i=w_s}^{w_e} (a_s(i) - E\{\overline{A_s}\})^2} \sqrt{\sum_{i=w_s}^{w_e} (b_s(i) - E\{\overline{B_s}\})^2}} \tag{3}$$

Where $[w_s, w_{s+1}, ..., w_e]$ is the joint support and $E\{\overline{A_s}\}$ and $E\{\overline{B_s}\}$ are the means of the corresponding series. The correlation coefficient is in the range of $-1 \leq corr \leq 1$, where the bounds indicating exact match up to a scaling factor, while 0 indicates no correlation.

Fig. 1 depicts the chart vectors of 6 different songs on each chart. The two Billboard charts are on a 1-100 scale, while the P2P chart is on a 1-2000 scale. The horizontal axis (x-axis) depicts the date measured in week numbers in 2007. The song titles and performing artists are written above each graph. Note that the lower parts of the graph represent higher position on the charts. Looking at Fig. 1, one can easily notice the high correlation, which is vivid not only in the general trend of the line, but also in minor trends and fluctuations.

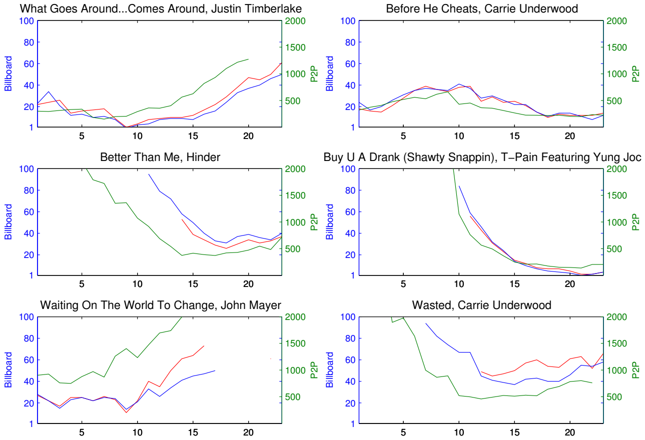


Figure 1. P2P Popularity Chart (green) vs. The Billboard Hot 100 (blue) and Billboard Digital Songs (red)

	Songs	avg. corr	median corr	avg. support
Hot 100	135	0.67	0.82	10.9
Digital Songs	113	0.67	0.8	10.7

Table II
Cross-Correlation: Our P2P chart vs. the Billboard charts

In all our measurements, we required songs to have a joint support of at least 4 weeks. Songs with a joint support of less than 4 weeks are mainly songs that ranked before or after our measurements, and had only a short "tail" inside our measurement period. Such songs poorly represent correlation of popularity trends over time.

First we measured the cross-correlation between the Billboard Hot 100 chart and the Billboard Digital Songs chart. According to (3), we measured the correlation coefficients of the 109 songs that had a joint support of 4 weeks or more. The average correlation coefficient was 0.88 while the median was 0.95, which indicates a very strong correlation. This high correlation between the two Billboard charts is somewhat expected. Let us now investigate the correlation between the the Billboard charts and our own P2P charts.

We measured the correlation coefficients of the 135 songs on the Billboard Hot 100, and the 113 songs on the Billboard Digital Songs that had a joint support of at least 4 weeks with the P2P chart. The results which are summarized in Table II, show that for the Hot 100 chart the average correlation coefficient was 0.67 while the median was 0.82, and for the Billboard Digital Songs, the average cross-correlation coefficient was 0.67, and the median was 0.8. These results indicate that songs on the Billboard charts are highly correlated with our independent P2P chart.

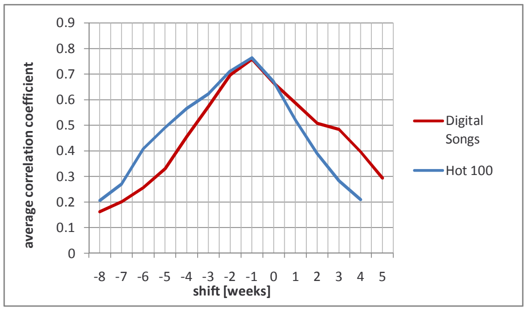


Figure 2. Cross-Correlation Coefficients vs. Time Shift

	Songs	avg. corr	median corr	avg. support
Hot 100	130	0.76	0.89	10.8
Digital Songs	108	0.76	0.88	10.7

Table III
Cross-Correlation: Our P2P chart vs. next week's Billboard charts

One might argue that the high correlation coefficients are the result of trend similarities of any time series of songs on charts. We thus measured the cross-correlation coefficient between the songs in the Billboard Hot 100 chart, and a random permutation (a different song) in the P2P chart. Of the 52 random matches which had a joint support of at least 4 weeks, the average joint support was 9.72 weeks, the average correlation coefficient was -0.006, and the median was 0.023, which negates the above hypothesis.

As mentioned is Section II, the Billboard charts were dated according to their release date. However, the data used to compile each chart, is collected during the 10 days before the chart is published. We were thus interested in the correlation coefficient between the P2P chart and the Billboard charts of the following week. In order to compensate for the delay caused by the Billboard's data collection process, we shifted the Billboard charts vectors backwards, and repeated our previous measurements. The results, which are summarized in Table III, show a substantial increase (0.06-0.08) from the previous measurements, which means that the vectors fit better with a one weeks time shift. Fig. 2 depicts the average correlation coefficients, as a function of the charts time shift. Clearly, minus one is the optimal time shift. The implication of this finding is obvious: P2P popularity charts can be used in order to predict trends on the Billboard charts.

B. Ranking Time Shift

Correlation does not reveal all the relations between P2P and Billboard. Here we look at the time difference between the week a song reaches its peak ranking on the P2P charts and

Chart	Average Ratio	Median Ratio
Billboard Hot 100	1.59	1.06
Billboard Digital Songs	1.59	1.11
P2P Queries	1.68	1.02

Table IV
Ratio of Ascending and Descending in Billboard and P2P

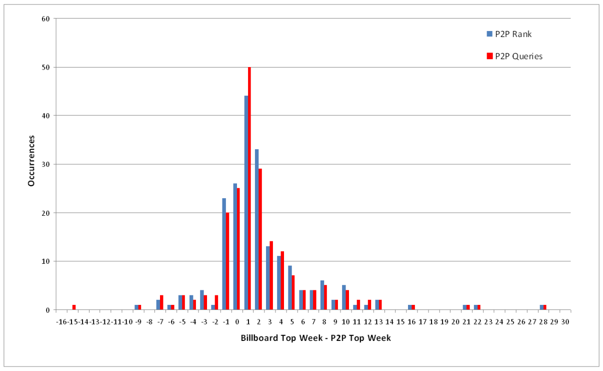


Figure 3. Billboard Hot 100 to P2P Top Rank Time Shift

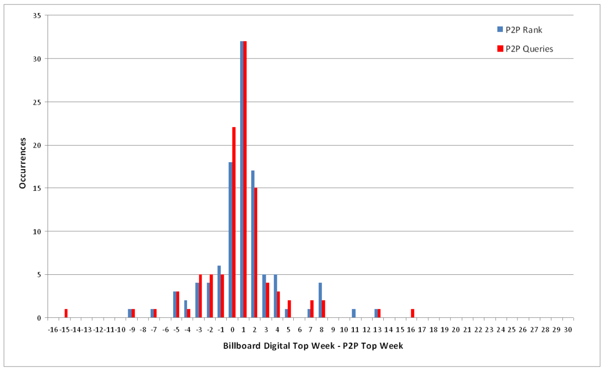


Figure 4. Billboard Digital Chart to P2P Top Rank Time Shift

the week it ranks highest on the Billboard. For the Hot 100 Billboard chart, 75% of the songs reach their highest rank and maximal P2P queries before the song reach its Billboard peak. On average, the time difference is 2.39 weeks. Figure 3 depicts a histogram of this information: the X-axis represents the time shift, the week number in which the song scored highest in Billboard minus the week number it scored highest in P2P. We measure time difference between the top rank on the P2P chart (blue bars), and the week we intercepted the maximum number of P2P queries (red bars). Each bar represents the number of collected occurrences. Clearly, most of the songs in the P2P network reach their peak and begin their decline before reaching their peak position on the Billboard. The dominant bin standing out above the rest represents the group of songs that reached their top Billboard rank exactly one week after reaching their peak on the P2P network.

In the Billboard Digital Songs chart, 63% of the songs reach their highest P2P rank before their Billboard's highest ranking. On average the time difference is 1.36 weeks, with a median of one. When the maximum number of queries is considered, 58% percent of the songs reached a maximum before reaching their Billboard peak. The mean time shift is 1.14, with a median od 1 week. Figure 4 depicts the histogram for the Digital songs chart. Again, the dominant bin is at a one week shift.

The above results are highly compatible with the results presented in Fig. 2: The Billboard's Hot 100 and the Digital Songs are highly correlated among themselves, and songs reach their peak about one week after reaching their peak on the P2P network. We further look at the gradient of songs ascending and descending the charts, or to be precise, the ratio between the two, normalizing scales. Table IV shows the ratios in the different chart. In all charts, songs climb slightly faster than they descend, and almost at the same rate in all cases.

IV. Ranking Prediction

Based on the ranking trends relations between the P2P network and the Billboard, we devised a prediction model for songs' Billboard top rank. The initial prediction is made on a song's debut week on the Billboard and is updated on weekly basis. The purpose is to identify hit songs that will reach high position on the Billboard, namely songs that will reach either the top 20 or the top 10. Our predictions are required to be causal, meaning we only use past and present information. Interchangeable attributes are of additional interest. We consider 2 types of attributes: the P2P popularity chart rank and the normalized number of P2P queries, meaning the portion of song queries out of the overall collected queries. The prediction models were built using familiar decision trees algorithms such as C4.5 [9], and BFTree [4], [11] implemented in the well established WEKA [13] data mining application.

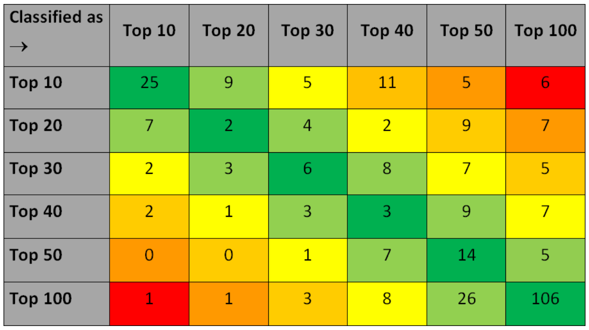
We consider several attributes that may contribute to such a prediction: We denote by NQ the total number of song's queries normalized by the total number of queries collected by our system, and NQ_{max} denotes the maximal value of NQ. Similarly, RQ will denote a song's rank in the P2P queries chart, with RQ_{max} being the maximal value of RQ. ∇RQ_n represents the change in a song's ranking in the n weeks prior to reaching its top rank. PRB denotes the predicted top Billboard rank and PRD the predicted digital song's rank.

A. Top Rank Numeric Predictor

At first we designed a predictor based on P2P information alone. Using Quinlan's M5 algorithm [9] we found a numeric predictor to a song's top rank in the Billboard Hot 100. The optimal prediction tree is a simple two leaf nodes decision tree based on the value of NQ. The decision criteria we got was: $NQ > 0.001$.
We denote by $LN1$ the linear predictor for $NQ <= 0.001$ and by $LN2$ the linear predictor for $NQ > 0.001$. The linear predictors for each node are the following:

- $PRB_{max\_LN1} = -73390.57 \cdot NQ_{max} + 58.96$
- $PRB_{max\_LN2} = -19164.22 \cdot NQ_{max} + 27.55$

The predictions were tested on 200 songs using a 10-fold cross validation. The correlation coefficient was 0.57 with an absolute mean error of 18.01 and a standard deviation of 22.3, which we consider to be high. We then classified the numeric results to the following discrete groups: Top 10, Top 20, Top 30, Top 40, Top 50, Top 100. This classification improves the precision of the prediction, as ranking resolution of ±5 is satisfactory.



Classified as →	Top 10	Top 20	Top 30	Top 40	Top 50	Top 100
Top 10	25	9	5	11	5	6
Top 20	7	2	4	2	9	7
Top 30	2	3	6	8	7	5
Top 40	2	1	3	3	9	7
Top 50	0	0	1	7	14	5
Top 100	1	1	3	8	26	106

Figure 5. Top Rank Confusion Table

Figure 5 depicts the confusion matrix of the discrete prediction. Rows represent the actual top rank of a song in the Billboard, while columns represent the predicted rank. In the Top 10 group, there is a 67.5% exact match, but if we allow the prediction to miss by one adjacent group, then for the Top 10 prediction we get a 86.4% precision and for Top 20 a 87.5% precision.

Looking on the entire group of songs predicted to reach ranks 1-20, we have 81.1% precision. Namely, out of the 53 songs we predicted to be in that group, only 10 failed to reach the top 20. Of these 10 songs who failed to reach the top 20, 5 reached the top 30 group. For the Top 30 and Top 40 groups,the precision achieved is 84% and 81.5%, respectively. The algorithm is also effective in predicting flops: It predicted songs that fail to reach the top 50 with a 78% accuracy.

On the Digital Songs chart the M5 algorithm was reduced to a simple one rule linear predictor (no branches):
$PRD_{max} = -32675.84 \cdot NQ_{max} + 37.93$
The absolute error was 12.87 and the standard deviation 15.57. This predictor is especially successful in identifying hit singles, with a 91% precision for Top 10 songs. However, it is a bit "optimistic" as it failed to identify any of the songs that ranked below top 40.

The initial prediction is done on a song's Billboard debut week. Therefore a song's Billboard debut rank is a valid attribute that might have additional explanatory information for the M5 algorithm. We thus repeat the above experiment, adding the song's debut rank as an input for the M5 algorithm. For the Hot 100 chart, we received the following regression rule:
$PRB_{max} = -0.43927 \cdot NQ_{max} + 0.4138 \cdot BB_1 + 21.77$
where BB_1 denotes a song's Billboard debut rank. This prediction has an absolute error of 15.37, a standard deviation of 19.30 and correlation coefficient 0.7. Predictions of Top 10 songs have a 78% accuracy. Top 20 predictions have a 84.3% accuracy.

P2P ranking is an interchangeable attribute for this prediction. Using RQ instead of NQ leads to a prediction with 2 leaf nodes, an absolute error of 14.7, standard deviation 18.54 and correlation coefficient 0.73. Though this predictor correctly predicts 90% of Top 10 songs, its set of predicted Top 10 songs is very limited - only 10 songs, compared to 28 using the first predictor.

For Digital top 100, the prediction is also improved:
$PRD_{max} = -24439 \cdot NQ_{max} + 0.4565 \cdot BB_1 + 13.60$
This prediction has an absolute error of 10.13, with standard deviation of 12.97 and a 0.69 correlation coefficient. Top 10 prediction has an impressive 96.5% accuracy. The one song that failed to reach the Top 10, ended up in the Top 20. For the Top 20 prediction, the precision was accurate on 89.2% of the songs.

B. Top Rank Classifier

The previous predictors were based on the numeric M5 algorithm. The classification to discrete groups was conducted after the numeric prediction. Here we used Quinlan's C4.5 classifier [9] for direct classification of rank groups. For each song, the classifier needs to decide whether it will reach the top 10 or not, and whether it will reach the Top 20 or not (Yes/ No). We start with the Hot 100 chart. Interestingly, the algorithm failed to distinguish between the top two groups, and all the songs that were predicted to reach the Top 20, were also predicted to reach the Top 10. We received a simple decision rule: $RQ_{max} < 60$
Namely, a song will reach the top 20 (and Top 10), if its P2P rank is above 60 (lower in absolute numbers). Since the algorithm fails to distinguish between the Top 10 and the Top 20, we analyze Top 20 prediction alone. In that case, the overall precision is 82%, with 87.8% for detecting songs that do enter the Top 20, and 78.9% for detecting songs that do not. On average, songs that do pass this threshold, do it 2.83 before reaching the Billboard Top 20.

Adding Billboard debut rank slightly improves the results: For Top 10 prediction, the algorithm created a 4 leaf nodes tree, as shown in Figure 6. The precision here was 85.6%, with 91.4% accuracy in identifying songs that do not make the Top 10, and 66.7% in identifying songs that do.

The decision tree for Top 20 prediction is a 3 leaf nodes tree: A song will enter the Top 20 either if it is ranked in P2P chart above 60 or if its debut Billboard rank was above 24. Here, we had a 86% overall precision, but with 88.7% precision for detecting Top 20 songs and 84.6% for detecting songs that fail to reach the Top 20. A prediction based on the same algorithm using the Billboard debut rank alone, gives precision rates lower than 80%. We thus conclude that our P2P chart does add additional explanatory information beyond that of the debut rank.

Results were less accurate in the Digital Songs chart. For Top 10 predictions, the overall accuracy is only 73.5%, with just 50% for detecting Top 20 hits, and 81.5% for detecting songs that do not make it to the Top 20. Additional information such as NQ, ∇RQ_n, or BB_1 failed to improve accuracy. For the Top 20 prediction, we received again a simple single rule predictor: $RQ_{max} < 294$
Namely, a song will reach the Top 20 if its Top P2P rank is above 294. The overall precision for such a prediction is 75.8%. 68.5% of the songs that were predicted to reach the Top 20 actually reached it, however 40% of those who failed

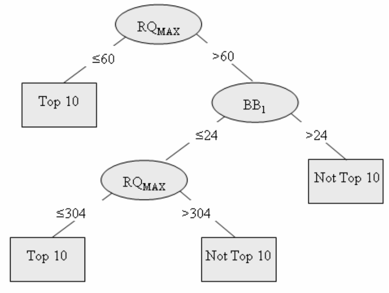


Figure 6. Top 10 Decision Tree

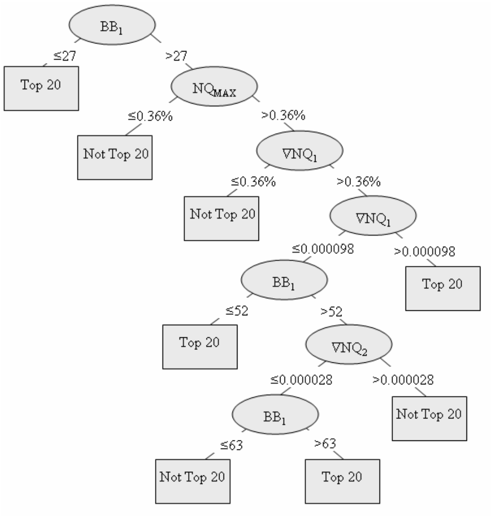


Figure 7. Digital Top 20 Decision Tree

to reach the Top 20, did reach the Top 30. Songs reach P2P rank above 294 an average of 4.5 weeks before they enter the Digital songs chart Top 20. A more complex decision tree shown in Figure 7 can improve the overall precision to 77.6% and increase the amount of songs detected. Here, we correctly identified 33 of the 46 songs that entered Top 20 during the examined period.

V. Conclusions

In this paper we have explored the relations between P2P and Billboard charts, showing a strong correlation between P2P queries and both Billboard Hot 100 and Digital Songs charts. It is discussed how P2P queries reach their peak at the same time as a song reaches it highest Billboard ranking, thus showing that P2P downloads and music sales are closely tied together, with little to no time gap. Yet, the P2P information is available a week before the Billboard charts are released. We suggest several novel prediction models of a song's success in the Billboard based on P2P queries and P2P popularity chart ranking. We manage to predict the success of a song in the Billboard Hot 100 with over 86% precision, and in Billboard Digital Songs with over 89% accuracy. The discovery of a hit song can be done 2 to 3 weeks ahead of the Billboard chart release based on P2P queries information. In our future work we intend to focus more on understanding the trends between Billboard and P2P networks, as well as developing prediction models for songs lifetime in the Billboard and predicting song change of ranking in the Billboard per week.

VI. Acknowledgements

We would like to thank Oded Maimon for his interesting insights.

This research was supported in part by a grant from the Israel Science Foundation (ISF) center of excellence program (grant number 8008/03).

References

[1] S. Bhattacharjee, R. Gopal, K. Lertwachara, and J. R. Marsden. Whatever happened to payola? an empirical analysis of online music sharing. *Decis. Support Syst.*, 42(1):104–120, 2006.

[2] S. Bhattacharjee, R. D. Gopal, K. Lertwachara, and J. R. Marsden. Using p2p sharing activity to improve business decision making: proof of concept for estimating product life-cycle. *Electronic Commerce Research and Applications*, 4(1):14–20, 2005.

[3] E. T. Bradlow and P. S. Fader. A bayesian lifetime model for the "hot 100" billboard songs. *Journal of the American Statistical Association*, 96:368–381, 2001.

[4] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression : A statistical view of boosting. *Annals of statistics*, 28(2):337–407, 2000.

[5] A. S. Gish, Y. Shavitt, and T. Tankel. Geographical statistics and characteristics of p2p query strings. In *The 6th International Workshop on Peer-to-Peer Systems (IPTPS'07)*, Feb. 2007.

[6] A. Klemm, C. Lindemann, M. Vernon, and O. P. Waldhorst. Characterizing the query behavior in peer-to-peer file sharing systems. In *Internet Measurement Conference*, Taormina, Italy, Oct. 2004.

[7] N. Koenigstein and Y. Shavitt. Song ranking based on piracy in peer-to-peer networks. In *10th International Society for Music Information Retrieval Conference*, October 2009.

[8] N. Koenigstein, Y. Shavitt, and T. Tankel. Spotting out emerging artists using geo-aware analysis of p2p query strings. In *KDD '08: The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 937–945. ACM, 2008.

[9] J. R. Quinlan. Learning with continuous classes. pages 343–348, 1992.

[10] P. Resnikoff. Digital media desktop report, fourth quarter, 2007. Digital Music Research Group.

[11] H. Shi. Best-first decision tree learning. Master's thesis, University of Waikato, Hamilton, NZ, 2007. COMP594.

[12] K. Sripanidkulchai. The popularity of gnutella queries and its implications on scalability, Feb. 2001. Featured on O'Reilly's www.openp2p.com website.

[13] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2 edition, 2005.