Visualizing Song Metadata

Mayank Gulaty
National College of Ireland
x15031705
August 5, 2016

***Abstract*—In this world of digital music, music metadata is now even more important than it was before the age of iTunes and yet it is overlooked.The purpose of this project is to see the trends in the music based on music metadata like song duration, genre and title and present this information in an interactive way so that the end user can manipulate the variables and see the data he wants to see rather than the static visualization. To achieve this, various points have been taken into consideration regarding interactive principles.**

***Keywords*—Music, Genre, Data Visualization, Visual variables, Interactive Visualization**

I. Introduction

"If history were taught in the form of stories, it would never be forgotten" Rudyard Kipling.

This applies everywhere. Human mind needs a kind of connection or a story to remember a lot of details at once. E.g. If they are given a list of random 10 things, they wont probably remember all of them and definitely not in a correct order. But if they are given a connection between each word and makes a story with that words, then they are definitely going to remember all of the list and that too in the correct order. Data and stories together have an effect on both emotional and intellectual level. [1]

Human brain remembers new things by associating them with the things you already know. It keeps on associating till the time it reaches the stage for new things to become associated. Thats where, the role of data visualization comes to deliver the previous unknown information in the most effective way.

Visualization is not just about making graphs. Its more than that. Many factors have to be considered before making a good visualization including color, shape, user perception etc.

II. Background

"Music metadata is often an overlooked part"[2] but on the other hand it has so many applications because it is independent of the song itself. So even if something happens to a song, metadata will stay intact. With al of these usefulness of the metadata, some researchers took it even one step further. The researchers at University of Bristol, England developed a system [3] that can predict if the song will be a hit with an accuracy of 60 percent. They did it by analyzing the UK top 40 charts over the past 50 years. But they did it just by the properties of audio music like temp, loudness, duration etc. However, the success in music industry is not just about good music rather there are many other factors involved other than just music. E.g. the title of the song also plays a very important role.

O. Creanor is with the Department of School of Computing, National College of Ireland, Dublin, Ireland

A. Why are song titles important?

Song title is the first thing user sees and is a crucial factor for a song to become hit or flop.

Some people think that the lyrics are the most essential part of the song but see the below questions [4]:

- When a user buys a song online, what is the first thing they see?
- How are the songs and videos titled on online social media and e-commerce websites?
- How are the songs called in the radio?

The answers to all of the above questions is a song title. Your songs title can capture the fans imagination and cause them want to know more or not [4]

B. Observed Metrics

According to [15], Data Visualization is the effective communication of data. [16] says that the visualization should make the user think about the information and not the design and make a large dataset easy to understand. However, not all the data can be represented in a single visualization. [5] in his book talks about different types of data nominal and ordinal. Ordinal includes the numeric values and nominal includes the non-numeric values. This project uses many different types of data. Ordinal data types in this projects are Year which is a discrete data type and can have values between 1960 and 2010. Time, number of weeks and score points are also discrete variables. Whereas, genre falls under the category of nominal data type. It is categorical as it has a finite number of values to select from. Artist names and titles are also nominal data types under the category of arbitrary as there can be potentially infinite number of values for these two variables. The data is collected from the Whitburn Project [6] which has around 120 years of Billboard [7] chart information from 1890 2010. More details about the data will be discussed in methodology section.

C. Related Work

There have been previous works done with the whitburn project [6] but most of them are static visualizations where

user has no control over it. E.g. [8] did the visualization of the most used words in the title of the pop songs. The bad thing of this visualization is that that user cannot select the genre or the time duration.

There are many things to consider to build an effective visualization. Design principles do not work every time. e.g. [9] tried different methods to visualize musical genres including colors, avatars, fonts and icons using online surveys but they couldnt come up with one approach as there were different responses from different people. As mentioned in [10], different parts of the world chose different color for the musical genre. Likewise, there are many visual representations of representing a variable in a visualization like size, texture, shapes, orientation, color etc. Colors are excellent for labelling and categorization but poor for displaying shape, detail or space [11]. This paper choses color for representing the genres as other variables wont fit in the visualization. Color scheme would be chosen as mentioned in [10] and is in perfect harmony with color symbolism [12]. E.g. red color is chosen to represent rock music as both represent violent behavior, aggressiveness, energy etc. and green was chosen to represent country as both symbolizes grass, tree, plants etc.

For making the word cloud, size was used to represent the words.

According to [13], there are 7 categories of interaction in a visualization select, explore, reconfigure, encode, elaborate, filter and connect. Select deals with the techniques to mark something that user finds it interesting. Explore means to show the user things that are not visible in the current visualization. Reconfigure just changes the way data is shown in order to easily understand the data. Encode changes the way data is represented. Elaborate feature give the user power to choose how much information to display. Filter gives the freedom to explore the data based on some conditions and connect feature connects the different views to make the visualization cohesive. This project will mainly use elaborate and filter techniques.

III. Methodology

A. Project Management

The Gantt chart 1 below shows the outline of the project plan that was followed. Firstly, the data was identified and a project plan was made.There were 3 milestones during the eleven week project. After proper data inspection, key objectives were thought of that what variables to visualize and what to omit. After selecting all of that, data was prepared, cleaned and organized. Then the data was pre-processed or statistical analysis of the data was done to view some interesting insights. After that, various visual principles were thought of and the visualization was made after which interactivity was added. At the end report was made, and the video was shot and submitted. All the milestones were met in the way.

B. Data Cleaning

Data cleaning had several missing values or was not in the appropriate format to go on to visualize. E.g. time was in the format ”mm:ss” but had to be converted to ”ss” for it to be

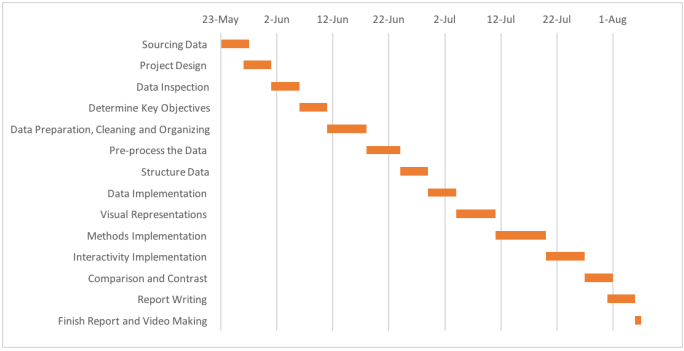


Fig. 1: Gantt Chart

processed and apply calculations. Genre value had more than 70 percent missing values but this column was important for the analysis so the values of the genre were collected via the musixmatch api [14] using the title and artist of the song. The data from 1960 onwards has been used for the project because the amount of missing data before 1960 was very large and it was not feasible to clean it due to time limitation of this project. This filtering was done through open refine by google [15] and rest of the cleaning was done in R. Open refine was used for the first step because the data file was too big and was not being loaded in R due to low configuration of the system. Some of the columns were not necessary and were removed.

C. Statistical Analysis

As mentioned by [5], statistical methods can provide insights between the dimensions that a user cant see it by naked eye. Therefore, before proceeding to randomly visualize the data, a correlation test was performed to see if there is a connection between the variables or one variable is affecting the other. For this project, Pearsons coefficient of correlation was used. The result is between -1 and +1 where the values close to the extremes show a strong correlation and the sign tells the polarity.As shown in 2, some of the insights were obvious like there is a strong positive correlation between score points and all the number of weeks variables. We can see that there is negative strong correlation between score points and high variable which seems to be ironical but it is because of the structure of the rank as small rank is better. Instead, it is also a positive correlation. One interesting observation was between the time of the song and the number of weeks in top 10 which was a moderately strong negative correlation which means if the duration of the song was long then the less number of weeks it will stay in top 10 charts.

As title is being discussed in the paper, a correlation test 3 was performed between the length of the title and the score points of the song and unexpectedly there was a moderately strong negative correlation which means that if the song title is long, the less popular the song is.

D. Design

The project was implemented using an iterative top-down approach. In this approach we first develop,design and analyze

Correlations

		ScorePoints	nWeeksChart	nWeeksChart Top40	nWeeksChart Top10	High	Time
ScorePoints	Pearson Correlation	1	.828**	.920**	.769**	-.816**	.170**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	22204	22204	22131	22131	22131	22204
nWeeksChart	Pearson Correlation	.828**	1	.778**	.507**	-.761**	.267**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	22204	22278	22170	22170	22170	22278
nWeeksChartTop 40	Pearson Correlation	.920**	.778**	1	.734**	-.812**	.135**
	Sig. (2-tailed)	.000	.u00		.000	.000	.000
	N	22131	22170	22170	22170	22170	22170
nWeeksChartTop 10	Pearson Correlation	.769**	.507**	.734**	1	-.538**	-.483**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	22131	22170	22170	22170	22170	22170
High	Pearson Correlation	-.816**	-.761**	-.812**	-.538**	1	-.084**
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	22131	22170	22170	22170	22170	22170
Time	Pearson Correlation	.170**	.267**	.135**	-.483**	-.084**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	22204	22278	22170	22170	22170	22278

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 2: Correlation 1

		ScorePoints	TitleLength
ScorePoints	Pearson Correlation	1	-.365**
	Sig. (2-tailed)		.000
	N	22204	22204
TitleLength	Pearson Correlation	-.365**	1
	Sig. (2-tailed)	.000	
	N	22204	22278

**. Correlation is significant at the 0.01 level (2-tailed).

Fig. 3: Correlation 2

and if its good to go, its released otherwise back to design process. Each design is iterated until almost all the flaws are gone. Audience here is a very crucial factor for whom the visualization is being made. This project focusses on people wanting to know some interesting insights from the song metadata. The visualizations are easy to understand and don't require much domain knowledge. After that data was categorized and visual symbols were thought of.

E. Implementation

Visualizations were implemented using Tableau and R. Tableau was used due to its simplistic model and easy to create visualizations. However more complex visualizations couldnt be made through Tableau, thats why R was used due to the power it gives. Shiny package was used to create the interactivity in the visualization. The visualizations are simple and self-explanatory for the end-user to understand.

IV. Results and Conclusion

A. Title

As mentioned above, the first visualization 4 done was between the title length and the score points and as stated above there seem to be a relation between the two. As we can see from the figure 4, the score points keep on decreasing as the length of the title is increasing. The visualization is interactive and user can filter on either genre or time duration. Below is an example filtered on "Country" and "Pop" genre during the year 1960-2010. We can also see that the best length for a title should be between 10-20 characters.

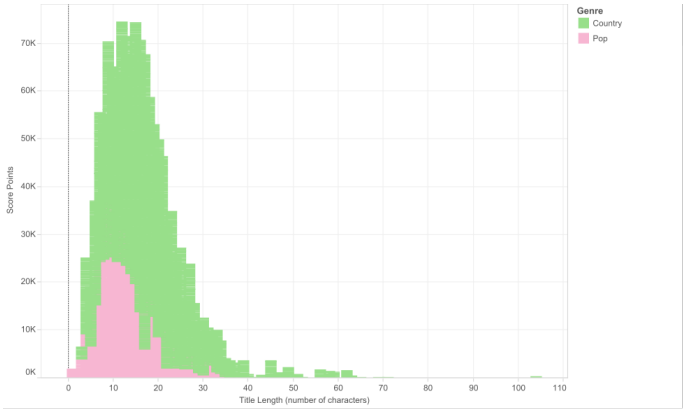


Fig. 4: Title length vs Score Points

The second visualization 5 is showing the most popular words in the title. User has control for selecting a particular year or range of years and can select the frequency of words in the word cloud. In the image below, the range is set from 1990-2000 and the most popular word we can see is love. An interesting observation from this visualization is that love was the most used word when selecting any filter. This visualization was very complex to make in tableau and was made in R using shiny package [16] which is useful for making web applications in R.

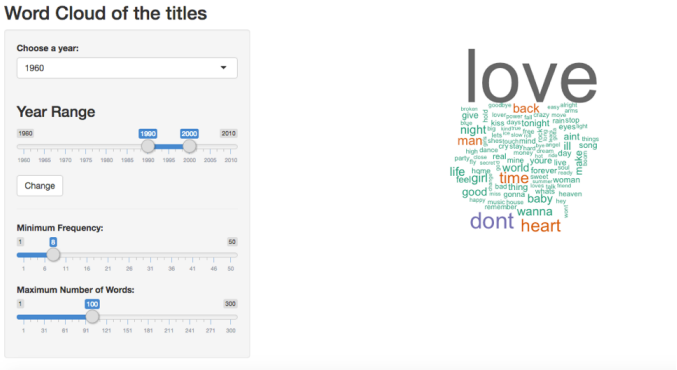


Fig. 5: Word Cloud

B. Time

As seen from the above correlation table 3 , we saw that duration of the song had an effect on the number of weeks, song will last in top 10 charts. So, the next visualization will be between the duration of the songs and number of weeks in top 10 6. As expected, there seem to be a pattern between these two variables. We can see that the songs with duration

greater than 280 seconds normally dont list in top 10 even for 1 week except 2 or 3 exceptions. The below example is filtered on country genre and as above, user has the power to select genre or time duration.

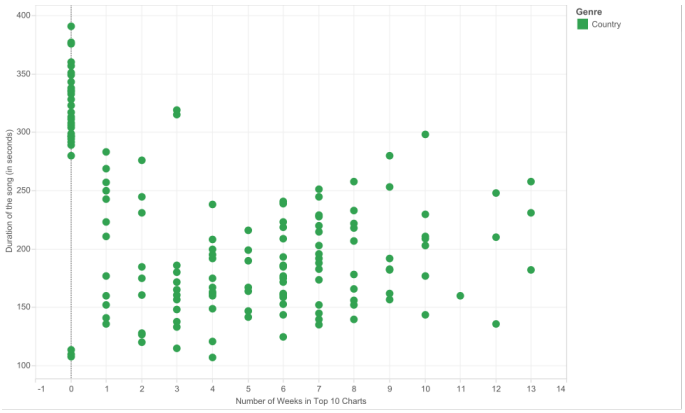
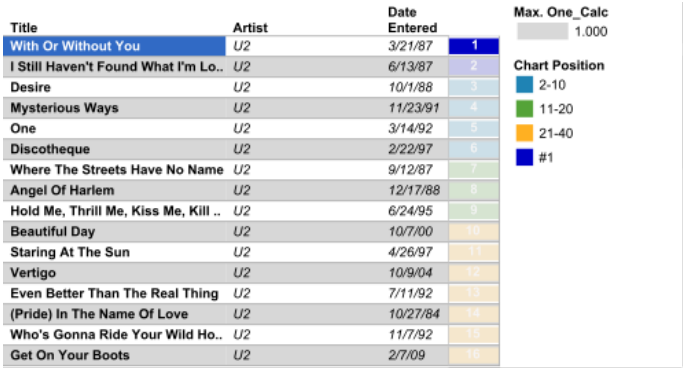


Fig. 6: Time vs Number of weeks in Top 10

The next visualization 7 is a basic one to see the trend of the minimum, average and maximum duration of a song and how it compares with the other genres. The figure below shows the comparison between pop and country genre over the period of 50 years from 1960 to 2010. We can see the trend from the below output that the duration of country songs was greater than the pop songs but the trend changed from 2014 when there was a sharp increase in the duration of pop songs.

Fig. 7: Duration vs Year

Last visualization is the one giving the user most control. This visualization consists of 3 parts. Firstly, the user can search a title directly or search an artist and then select a title. The second view shows the songs of the artist in a chronological order giving additional information after hovering the mouse over it. The third part shows the selected songs performance in a more detailed way using the elaborate interaction technique mentioned in by [13]. In the example below artist U2 was searched and it gave the first output with all the songs 8 . After we select a song, in this case With or without you, it will highlight the song in the second tab 9 and shows the detailed performance in the third output 10 .



Title	Artist	Date Entered	
With Or Without You	U2	3/21/87	1
I Still Haven't Found What I'm Lo..	U2	6/13/87	2
Desire	U2	10/1/88	3
Mysterious Ways	U2	11/23/91	4
One	U2	3/14/92	5
Discotheque	U2	2/22/97	6
Where The Streets Have No Name	U2	9/12/87	7
Angel Of Harlem	U2	12/17/88	8
Hold Me, Thrill Me, Kiss Me, Kill ..	U2	6/24/95	9
Beautiful Day	U2	10/7/00	10
Staring At The Sun	U2	4/26/97	11
Vertigo	U2	10/9/04	12
Even Better Than The Real Thing	U2	7/11/92	13
(Pride) In The Name Of Love	U2	10/27/84	14
Who's Gonna Ride Your Wild Ho..	U2	11/7/92	15
Get On Your Boots	U2	2/7/09	16

Fig. 8: Artist Songs

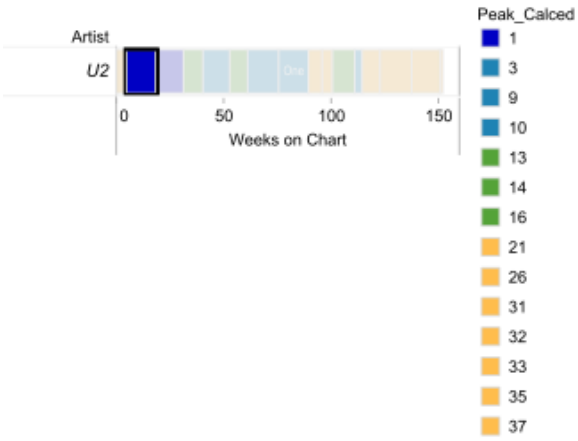


Fig. 9: Artist Peak

V. Future Work

The information can also be presented in a more sophisticated way as shown below but due to time limitation was not possible. The points on the circle show the depict the week number and the polygon represents a song 11a . The artists image is at the middle of the circle. The closer the vertices of the polygon are to the points of the circle, the higher the rank is in that week. More corners of the polygon mean more weeks on top charts. The higher the area of the polygon, the more popular the song is. The visualization is for an artist, so there will be multiple polygons in one graph depicting different songs by which the user can compare the performance of the songs. A more detailed graph of the song will be shown in a different graph after clicking on a polygon as shown in the second figure 11b .

References

[1] J. Aaker, "Jennifer Aaker: The Power of Story," Sep. 2013. [Online]. Available: https://futureofstorytelling.org/video/persuasion-and-the-power-of-story/

[2] "4 Reasons Why Metadata in Songs is Important." [Online]. Available: http://mp3.about.com/od/musicformats/fl/Song-Tags-The-Importance-of-Metadata-in-Music-Files.htm

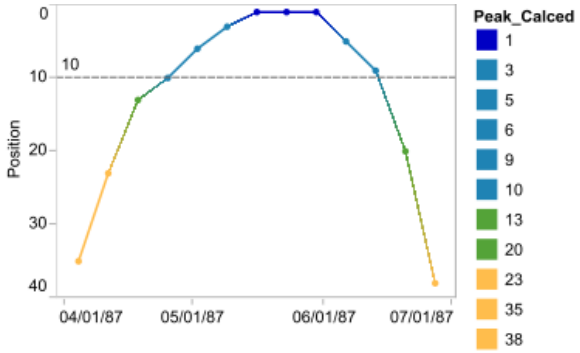


Fig. 10: Detailed song chart information

(a) Proposed Visualization 1 (b) Proposed Visualization 2

Fig. 11: Caption for this figure with two images

[3] "SCOREAHIT." [Online]. Available: http://scoreahit.com/

[4] M. Wise, "The Title of Your Song IS Important | Christian Band Help." [Online]. Available: http://christianbandhelp.com/title-of-your-song-is-important/

[5] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2010.

[6] A. Baio, "The Whitburn Project: 120 Years of Music Chart History." [Online]. Available: http://waxy.org/2008/05/the_whitburn_project/

[7] "Billboard - Music Charts, News, Photos & Video." [Online]. Available: http://www.billboard.com/

[8] T. Whitwell, "Music Thing: The 100 most clichd clichs in pop song titles." [Online]. Available: http://musicthing.blogspot.ie/2008/05/100-greatest-ever-cliches-in-pop-song.html

[9] J. Holm and H. Siirtola, "A comparison of methods for visualizing musical genres," in *2012 16th International Conference on Information Visualisation*. IEEE, 2012, pp. 636–645.

[10] J. Holm, A. Aaltonen, and H. Siirtola, "Associating colours with musical genres," *Journal of New Music Research*, vol. 38, no. 1, pp. 87–100, 2009.

[11] C. Ware, *Information visualization: perception for design*. Elsevier, 2012.

[12] J. Morton, *A guide to color symbolism*. Colorcom, 1997, vol. 28.

[13] J. S. Yi, Y. ah Kang, J. Stasko, and J. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE transactions on visualization and computer graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[14] "Musixmatch API." [Online]. Available: https://developer.musixmatch.com/

[15] "OpenRefine." [Online]. Available: http://openrefine.org/

[16] RStudio, Inc, *Easy web applications in R.*, 2013, uRL: http://www.rstudio.com/shiny/.