



Analysis of a Social Data Visualization Web Site: YouTube

Najla Aljabr¹, Ghadeer Alkalthm², Rawya bin abdulrhman³ and Muneera Alhabdan⁴

Department of Management Information Systems, Imam Abdulrahman Bin Faisal University (IAU),
Dammam, Kingdom of Saudi Arabia

ABSTRACT

This research discusses and analyzes the idea of data visualization, its characteristics, important elements and the concept behind it. The ideal data visualization model must be visually appealing, expandable and accessible, also develop the accurate information. Many challenges found that face analysts such as the complexity of data, lack of knowledge and over-relying from the users. The research will specifically evaluate results of YouTube using tools like Tableau and Rapid Miner, focusing in the most famous channels in different industries, helping companies to select a target and choose the right segment where to view their advertisements, which results in an increase in profit in the long term. After selecting YouTube data for eight different industries in Saudi Arabia, Top ten channels were chosen based on views and subscribes, one more step is to compare the numbers of videos uploaded and the number of subscribers using the advanced tools mentioned above. In Rapid miner, the method used for analysis is the K-Means algorithm which searches for k groups, cluster and calculate their correlation coefficient, then determine the results of data evaluation process.

Keywords: *YouTube, Data Visualization, Data Mining, Clustering, Decision Tree.*

1 INTRODUCTION

Data visualization, it's simply the appearance of data in graphical format, processing a huge amount of data in the fastest speed, it also could be interactive which means creating charts and graphs in more details and interactively deal with the data. This technique gives the ability to decisions makers to see the results of analysis visually to make it easier to understand or determine new methods. Nowadays, data visualization has a wider meaning involving many fields such as science and art, that will affect the business aspects in the future, and it's a good investment in the future of big data. Because of the way human mind process information, showing complex data in a graphical way will make it easier and faster to understand compared to reports or written documents.

Data Visualization can increase customer satisfaction, by identifying the spots that need improvement or where are the elements that have an impact on consumer's behavior and allow to predict sales volume. It is promoting creativity and it's going to change business analyst's way of dealing with data, they're expected to respond to problems faster and look at data in more different way.

Data Visualization tells you the story untold, storytelling is where social media and data visualization representation really crosses. Records recommend that social media is the sole purpose for the 90% data generated during recent years. This surely requires the need to visualize data. YouTube is one of the social media website that visualized data into geographic viewership that provides number of views, number of subscribers in every country even the time in hours or days that represent in graphs.

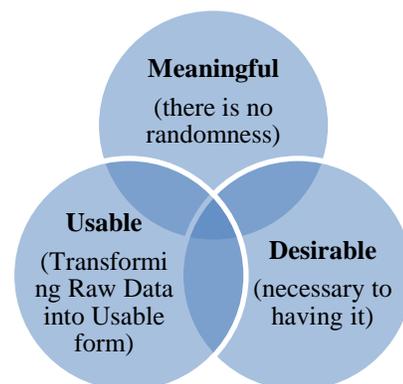


Fig. 1. Excellent data visualization

2 DATA VISUALIZATIONS CHARACTERISTICS

Data visualizations appear in multiple forms and sizes, but all have major characteristics that help arrange and to produce information with important insights. overall, a good data visualization model should be:

Visually appealing: The Visual appeal is what meets the eye, the colors, shapes, style and fonts. the concept behind the data visual appealing is to be an inspiration for the spectator. The new visualization tools and technology have raised the level of efficacy and usability. On the other hand, If the data visualization was developed with poor visual appealing and old technology, it will be a waste of time and effort. Moreover, the traditional data visualizations tools are becoming unable in the analysis on these days. [1]

Expandable: As the volumes of information available for analysis is increasing and the ways of information dispersion across the globe continue to diversify, the scalability of data visualizations is becoming a critical factor. Data visualizations scalability is a concept that emphasizes the capability of information to handle growth in the amount of data and users.

Accessible: Accessible data visualizations design includes allowing for flexibility, being simple and minimizing errors, also the people from different devices can access the visualization, whether it is PC monitor, mobile device, and the high-resolution monitor. [2]

Enables development and gives the right information: The data visualizations build on, updates, and extends previous efforts that communicate interesting information with both style and substance.

3 DATA VISUALIZATION AND CHALLENGES

The main challenge that data visualization developers face is the need to assure doing it in a reasonable, logical, and acceptable way. It builds a new dynamic, where the data overlaid needs to be obvious, brief and not complex or distracting. [3]

Data complexity

The non-oversimplification of data is one of the biggest problems of visualization, when it's cannot simplify them to more basic and understandable way. and the bigger the data, the more effort needed to query it. Furthermore, the complex data visualizations are more difficult to prepare and analyze rather than simple data, and often require a different set of business intelligence tools to do.

Overreliance on visuals

This problem faces the user more than specialists and developers, the person could easily start over-relying on it. However, trying to take millions of data as visuals could lead to unfounded conclusions, or could completely change the required.

Knowledge gap

The lack of knowledge in using data visualization, due to the academics do not get formal training in data visualization, as a result, researchers continue to rely on correlation tables and simple bar and plot charts to display data. [4]

4 RELATED WORK

The research is about the conversational behavior and social attention in YouTube and is a new way to analyze the video blogs (vlogs) by introducing a new research domain in the social media, and was named the automatic analysis of human behavior in conversational vlogs. Briefly, the aim of this research is to understand the processes involved in this social media type, based on verbal and nonverbal channels. The work contribution consists four parts. The first part is casting the vlog as a novel research domain in social media, comparing it to two type analysis, the text-based social media analysis and the face-to-face interaction analysis. Second, via the automated study of vlogging, bringing together the nonverbal behavior analysis and social media analysis. Third, to characterizing the vlogger behavior, by extracting audio, visual and multimodal nonverbal cues which are motivated by social-psychology. Finally, the research was done on a sample of 2,200 vlogs from YouTube, stating with hypothesize, then show and explain the specific details. [5]

The author research shows that comment spamming in YouTube has become a common phenomenon and there is a strong need to. they work in a method to automatically detect comment spammer in YouTube forums. The proposed technique is based on mining comment activity log of a user and extracting patterns indicating spam behavior. Results show that both campaigns are managed by a single spammer who controls a number of spam bot user accounts.[6]

5 METHODOLOGY

It is Impossible to analyze and evaluate the activity occurring with 1.2 million videos uploading every day, the traditional analysis techniques like (EXCEL and SQL database) are not enough. So, we use in this project some advanced tools like:

Tableau and Rapid miner, to concentrate on how the data generated from YouTube can be utilized by various famous channel in different industries, to help the companies to make targeted and informed decisions about what videos they can put an advertising product on it, to increase the awareness and profits of the company. Therefore, we are collecting YouTube Data for eight industries in Saudi Arabia for visualizations analysis.

- Brands
- Celebrities
- Community
- Entertainment
- Media
- Place
- Society
- Sport

For each industry, we are selecting the top ten channels and account according to the most uploaded video views and subscribers for the year 2017.

- Number of uploaded video views
- Number of subscribers

After that we are comparing the number of videos uploaded and the number of subscribers for each industry by using Tableau and Rapid miner program. Totally, we collected over 8 industries with 74 top data channels name.

6 RESULTS AND DISCUSSION

Table 1: The number of data visualizations by industry

Type	Total Top 10 Channel subscribers	Total Top 10 Channel views	Number of Channel	Name of top channel in each industry
Brand	1850317	1190987191	10	STC
Celebrity	5262297	1106082444	10	ARAM TV
Community	8557440	1643624281	10	Saudi Gamer
Entertain	311619716	6465773833	10	mmoshaya
Media	3034172	1305213670	10	Knary Alfdaaya
Place	51589	21782448	10	alhasa municipality

Society	458134	118358481	10	Ministry of Health
---------	--------	-----------	----	--------------------

A. Data Analysis in Rapid Miner

- Clustering:

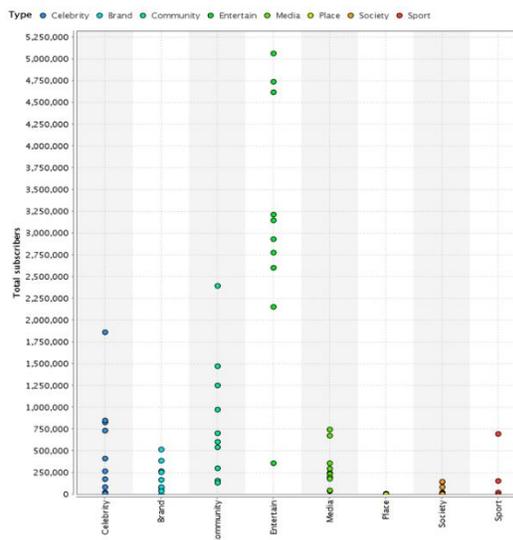


Fig. 2. the statistic presents the most viewed YouTube industry in KSA, sorted by yearly subscribers.

For this analysis, we are using the K-Means algorithm (which are clustering finds groups of data which are somehow equal.). This algorithm searches for the k groups, aggregated all the data and calculated their correlation coefficient. First, we change the type of attribute from non-numeric attributes to a numeric type, then we use the simple validation by splitting up the data relatively into a training set and test set by using 0.4 in a split ratio. After that we transform data by applying a preprocessing model. Finally takes a cluster model as derived index from the number of clusters by using the formula $1 - (k / n)$ with k. From the above graphs, we can see that as the number of subscribers in Entertainment industry is upper more the any industry, at the same time the place industry is the lower one.

- Repels Missing values (normalize)

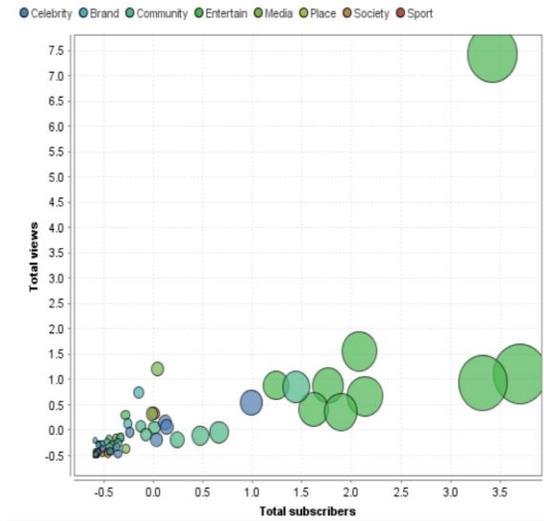


Fig. 3. the chart presents normalized data.

This figure shows the analysis that obtained by the rapid miner. In this analysis, we used the replace missing value first to replace the minimum, maximum or average value of Attribute. Zero can also be used to replace missing values. Any replenishment value can also be specified as a replacement of missing values. After that, we normalized that data which is used to scale values, so they fit in a specific range. Adjusting the value range is very important when dealing with Attributes of different units and scales. After normalization, we see the Entertainment industry is the most views and subscribers than others.

- Decision Tree

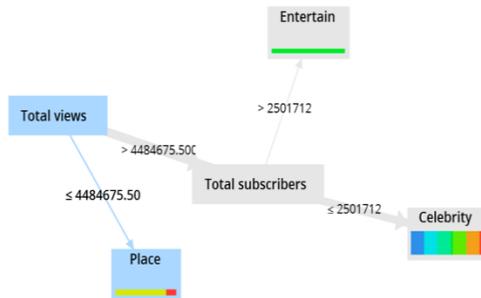


Fig. 4. the figure presents the most viewed and subscribers on YouTube at the decision tree

This figure, shows the decision tree which is a tree-like collection of nodes intended to create a decision on values affiliation to a class or an estimate of a numerical target value. Each node represents a splitting rule for one specific Attribute.

For classification, this rule separates values belonging to different classes, for regression it separates them in order to reduce the error in an optimal way for the selected parameter *criterion*. What we generated from this analysis is the total viewed of place is less than 4484675.50 on the other hand, the total viewed of celebrity and entertainment is upper than 4484675.50. Moreover, about the total subscribers the most subscribe is the entertainment more than celebrity.

B. Data Analysis in Tableau

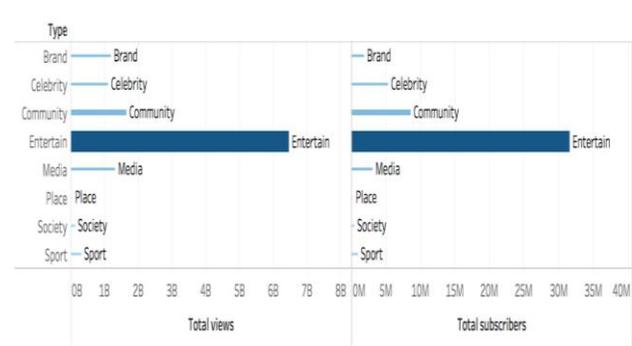


Fig. 5. Indicates the number of subscribers to the number of views in all uploaded videos

This figure shows how we obtained the number of subscribers and the number of uploaded video views in each channel has in each industry that was chosen from the eight categories. After we get these numbers for users in all the eight industries, results were that Entertainment industry is the highest among others, second is the Community industry in both sides, Media have the highest views, but the Celebrity industry beat it with number of subscribers. The industry of Brands has the subscribers to be less than total views, after it comes Sport, Society, and finally Place appears to be at the end of the list.

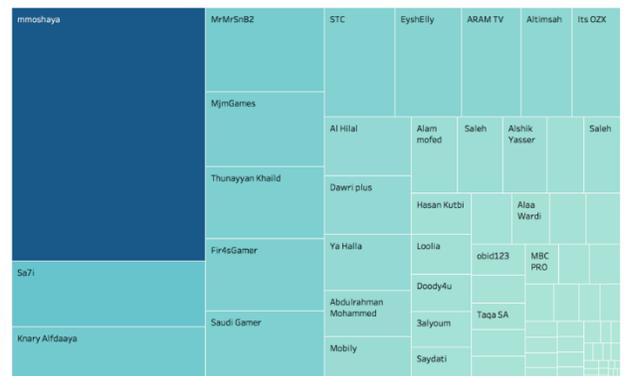


Fig. 6. Indicates the name of channels that got the highest rate of views and subscribes.

On this figure, it shows the names of channels that got highest rate of total views and total subscribers, those accounts are valuable for the company's advertisements and marketing, it also takes the demographics of views into consideration and based on which segment the companies want, who are the targets and what's their common behavior, they decide on which channel industry they want then advertise inside the video. Top channels founded to be "mmoshaya" in first place with more than 4.8 million subscribers, then it comes "Sa7i" in second place with more than 3 million subscribers.

7 CONCLUSION

In this paper, we have presented our preliminary visualization analysis of YouTube channel in KSA, to help the commercial company choose the right decision about the promoting their products on YouTube in front of a massive potential audience for maximum impact. Many huge companies that have implemented big data visualization are realizing significant competitive advantage compared to other company with no big data efforts. In this paper, we intended to construe the YouTube big data and come up with considerable insights which cannot be determined otherwise, by using the analysis technique which is Rapid miner and Tableau. One of the output results of data analysis shows that the YouTube viewers are interested in Entertainment industry more than any other kind. In the future, we could include extending the YouTube data analysis by using a comments analysis can be conducted to understand the attitude of people towards the specific video.

8 REFERENCES

- [1] "What Makes Good Data Visualization? – Dummies". [Online]. Available: <https://www.dummies.com/how-to/content/what-makes-good-data-visualizaiton.html>.
- [2] "The Top Trends In Data Visualization for 2018". [Online]. Available: <https://carto.com/blog/top-trends-data-visualization-2018/>.
- [3] "The 5 Biggest Challenges Facing Data Visualization | Articles | Chief Data Officer". [Online]. Available: <https://channels.theinnovationenterprise.com/articles/the-5-biggest-challenges-facing-data-visualization>.

- [4] "Data Visualization in Market Research", Research World, vol. 2014, no.47, pp. 10-19, 2014.
- [5] D. GATICA-PEREZ and J. BIEL, "VlogSense: Conversational Behavior and Social Attention in YouTube", Publications.idiap.ch, 2018. [Online]. Available: http://publications.idiap.ch/downloads/papers/2011/Biel_TOMCCAP_2011.pdf. [Accessed: 12- Apr- 2018].
- [6] A. Sureka, "Mining User Comment Activity for Detecting Forum Spammers in YouTube", Ir.ii.uam.es, 2018. [Online]. Available: http://ir.ii.uam.es/usewod2011/usewod2011_sureka.pdf. [Accessed: 12- Apr- 2018].