Viewed by Too Many or Viewed Too Little: Using Information Dissemination for Audience Segmentation

Bernard J. Jansen
Qatar Computing Research Institute, Qatar.
jjansen@acm.org

Jisun An
Qatar Computing Research Institute, Qatar.
jan@hbku.edu.qa;

Soon-Gyo Jung
Qatar Computing Research Institute, Qatar.
sjung@hbku.edu.qa

Haewoon Kwak
Qatar Computing Research Institute, Qatar.
hkwak@hbku.edu.qa

Joni Salminen
Qatar Computing Research Institute, Qatar.
jsalminen@hbku.edu.qa

ABSTRACT

The identification of meaningful audience segments, such as groups of users, consumers, readers, audience, etc., has important applicability in a variety of domains, including for content publishing. In this research, we seek to develop a technique for determining both information dissemination and information discrimination of online content in order to isolate audience segments. The benefits of the technique include identification of the most impactful content for analysis. With 4,320 online videos from a major news organization, a set of audience attributes, and more than 58 million interactions from hundreds of thousands of users, we isolate the key pieces of content in terms of identifying audience segments that are both (a) least and most discriminating in terms of audience segments and (b) the least and most impactful. By empirical methods, we show that 25.3 percent of the videos are so widely disseminated (i.e., viewed by so many different segments) that they are non-discriminatory, while 29.7 percent of the videos are very discriminatory (i.e., can clearly identify one or more audience segments) but their impact is marginal, as the user base is small. Implications are that there are critical values that can be identified to isolate the set of both distinct and impactful content in a given data set of online content. We demonstrate the utility of this line of analysis by using the approach to identify critical cut-off values for dynamic persona generation.

KEYWORDS

Social Media Analytics, User Experience Research, Data-driven design, Data Science, User Analytics, Market Segmentation

INTRODUCTION

Understanding the audience is a critical task in many domains, from marketing to advertising to system design to publishing to website architecture. In fact, there has been substantial work in many areas in understanding audience information consumption patterns (Hendahewa & Shah, 2017; Tan, Na, & Ding, 2015), as this is often critical to success in publishing and other areas. However, understanding the audience is a misnomer, as in many situations there is not one audience but many audiences, referred to as audience segments. Identifying audience segments in many contexts is difficult, for reasons ranging from a lack of data to privacy concerns to isolating what data to use. It is the last difficulty that motivates our work. Determining what content to use to accurately distinguish audience segments can be quite challenging. Some content may be widely viewed, but so widely that it is non-discriminatory. Other content may be extremely discriminatory but the volume of views may be so small as to not be impactful. This dichotomy motivates our research, as we want to identify that content that is both discriminatory and impactful for ascertaining audience segments for future content creation.

For this research, we use the online videos from a major news organization, along with associated audience demographic attributes and views from a major social media platform, to identify the optimal values for content separation for identifying both distinct audience segments (the upper value) and impactful audience segments (the lower value). We use the concept of information dissemination, which in this context refers to the breadth of audience segments to which a piece of content has been distributed.

We lead off with a brief background section, introduce our data site and data collection, and then present our methods and results. We then show the utility of the analysis by using the information dissemination analysis for the dynamic generation of personas for audience data. We end with discussion, implications, and directions for our next stages of research.

80th Annual Meeting of the Association for Information Science & Technology, Washington, DC, VA | Oct. 27-Nov. 1, 2017

BACKGROUND

Audience segmentation (Mellor, 2006) is the process of dividing a group of people into homogenous subgroups, typically based on behaviors and demographics, grounded around some product, brand, advertisement, or message (Stern, 1994), with many factors affecting audience consumption (Chen et al., 2017; Fu & Sim, 2011), which are brands target audience or customers. Identification of audience segments has long been important in marketing and advertising (Smith, 1956), and it is increasingly important in the technology and content publishing areas. Identification of audience segments is typically driven towards understanding a subset of people's reactions, interactions, uses, etc., based on one or more key performance indicators (KPI), in order to achieve some goal or objective, such as increasing revenue (Ortiz-Cordova & Jansen, 2012).

For this research, we investigate the use of information dissemination (Song, Lin, Tseng, & Sun, 2005) to identify meaningful audience segments, in terms of audience size, for content from a major news corporation. Some content may be widely disseminated among many audience segments. While good in one sense, this makes this particular content less valuable in discerning distinct segments. Other information may be very useful for identifying audience segments; however, its impact may be minimal if the audience segment(s) do not represent significant percentages of the overall user base or content views.

In this research, we are interested in using specific pieces of content (Zhang, Zheng, & Tai-QuanPeng, 2017), in this case videos, in order to discern key audience segments, which can then be used for key marketing and other decisions (Jisun An, Cho, Kwak, Hassen, & Jansen, 2016), including the development of software systems. The issue is that some content can very popular, viewed by some in nearly all audience segments (Wang, Zhou, Jin, Fang, & Lee, 2017). In these cases, these pieces of content are not very useful for discerning audience segments. Conversely, there may be content that is very discriminative in that it is viewed by one or few audience segments. However, if these audience segments are small, the actual impact of the content is minimal. It is this situation that motivates our research, as personas can represent only the meaningful audience segments.

RESEARCH OBJECTIVE

Our objective is to develop metrics for identifying online content with which an audience interacts that both identify distinct audience segments (i.e., information discrimination) and impactful audience segments (i.e., information dissemination). These two concepts are used to create upper and lower bounds, isolating the data to be used for analysis or system development.

Such metrics would be valuable in a variety of venues including web analytics analysis and design of systems for audience analysis.

As part of this research, we implement our approach in the design of a system for dynamically generating personas, which is a technique used for understanding audience segments by use of a fictitious person.

DATA SITE AND COLLECTION

Using actual user data from AJ+, a major online media and mobile channel based in the United States, we validate our premise concerning the discriminatory value of information dissemination for audience segmentation. In the highly competitive online news industry, an understanding of audiences is notably important to both increase the consumption of digital content and also get relevant and noteworthy information to the readers that may be impacted by the news events.

Data Site: AJ+

AJ+ is natively digital content platform. AJ+ was designed from the ground up to serve news in the medium of the viewer, with no redirect to a website. AJ+ is based mainly on social platforms. Therefore, the digital content is specifically designed to be viewed in the Facebook Newsfeed, YouTube Channel, or Twitter Timeline depending on the readers who are most active (i.e., interacting with online content.

Data Collection: AJ+ YouTube Channel

For the data collection for the research reported here, we use the AJ+ YouTube Channel, reserving analysis for the Twitter and Facebook platforms for future work, as the technique is generalizable to any channel. The principal reason to focus on the YouTube channel is that the analytics platform gives the detailed statistics for every video, although we believe the approach transferable to other social media platforms as well. As an example of an AJ+ YouTube video, see Figure 1, noting the number of views.

***Figure 1**. Example of YouTube video from the AJ+ YouTube channel, with number of views*

Being quite robust, the YouTube analytics platform provides, for each piece of the AJ+ content collection, user profile attributes (e.g., gender, age, country location, and which site the user comes from), also at an aggregate level. We use these user and video attributes to explore if information dissemination can identify meaningful audience segments values based

on video content interaction and related demographics provided by YouTube.

One can access the data in the YouTube analytics platform by using the YouTube APIs[1]. The parameters we use for this research are listed below. There are various video KPI metrics; we focus in this research only on *viewCount* (the number of views).

For our data analysis, we collected 4,320 videos uploaded on June 13, 2014 to July 27, 2016, that had more 58 million user views. We should note that the data values for the YouTube channel is private and available only to the owner of the channel (i.e., AJ+ in this case), and thus not publicly accessible.

- **Audience Attributes**
  - *ageGroup* - YouTube viewers are classified into multiple age categories (13-17 years, 18-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, and 65 years and older), so there are seven possible age categories for a user.
  - *gender* - YouTube viewers are classified into either male or female, so there are two possible categories.
  - *country* - YouTube uses the two-letter ISO-3166-1 country code index to classify where viewers are from, with 249 current officially assigned country codes at the time of this study.

- **Video Attributes**
  - *viewCount* – YouTube provides the number of views per video.

METHODS

We now begin our approach for identifying the portion of the data set that is both distinct and has an impact. We first look at how widely the content is disseminated (although there may be a correlation between how widely (i.e., breadth) that a piece of content is has been viewed by members of audience segments and the number (i.e., volume) of views, this is not a necessary condition).

We define information dissemination, $I_{Dissemination}$ as:

$$I_{Dissemination} = \frac{Segments_{Information}}{Segements_{Total}}$$

where $Segments_{Information}$ is the number of audience segments that interact with a piece of content from the complete set of information items, and $Segments_{Total}$ is the number of audience segments that have interacted with ***any*** item of information from the set. We define $I_{Dissemination}$ for each piece of content in the collection.

We then examine how useful a piece of content is for identifying distinct audience segments. We define information discrimination, $I_{Discrimination}$ as:

$$\tau_L < I_{\text{Discrimination}} < \tau_U$$

where tau with a subscript L, τ_L, is a predefined lower limit and tau with a subscript U, τ_U, is a predefined upper limit of audience segments. Particularly, τ_L is determined by the number of content views. τ_U is determined by a percentage of the number of audience segments. Presently, τ_L and τ_U are heuristically determined (i.e., we use rules logically determined). We leave more rigorous analysis in calculating tau values for future work.

In this case, we are interested in an optimal balance between $I_{Dissemination}$ and $I_{Discrimination}$ for videos, although the approach could apply to any distributed information content for which one is interested in audience segmentation.

Videos that have a large $I_{Dissemination}$ would not be good candidates to identify unique audience segments, as these videos are available and can be watched by for many audience segments, which informs us of τ_U.

Videos that have a small $I_{Dissemination}$ might not have enough impact, as measured by a metric such as *viewCount*, to make the audience segment identification worthwhile, which informs us of τ_L.

With this information dissemination approach, we analyze our data set to determine appropriate $I_{Dissemination}$ and $I_{Discrimination}$, values to identify both distinct and meaningful audience segments.

RESULTS

Given our dataset, we define an audience segment as a unique combination of (*country*, *gender*, *ageGroup*). So, with two gender groups, seven age groups, and 249 counties, we have an upper limit of 3,486 audience segments. (i.e., 2 x 7 x 249). In actuality, our data set has 2,214 audience segments, as the data has users from 190 unique countries and not all age groups for each country are represented. Thus, in the rest of this work, we use 2,214 as the baseline for maximum audience segments ($Segments_{Total}$).

We then calculate $I_{Dissemination}$ for each of the 4,320 videos in our data set toward our goal of identifying meaningful audience segments. We first determine the number of unique audience segments ($Segments_{Information}$) for each video. The average audience segment per video is 200, and the median is 136. The maximum audience segment is 1,980, and the minimum is 10. The standard deviation is 208.45.

It is interesting to note that the video with the maximum number of audience segments has an $I_{Dissemination}$ of 0.8618, meaning that more than 86% of all audience segments viewed this video. Obviously, this video would not be a discriminative candidate in our dataset.

Conversely, the video(s) with the minimum number of audience segments has an $I_{Dissemination}$ of 0.0045, meaning that 0.45% of audience segments viewed this video. So, the three

[1] https://developers.google.com/youtube/analytics/

videos with an $I_{Dissemination}$ of 0.0045 have good discriminative power. However, when we examine the *viewCounts* of each of these videos, in total, they represent 0.0012% of total views, which is not impactful. Therefore, these audience segments are not meaningful.

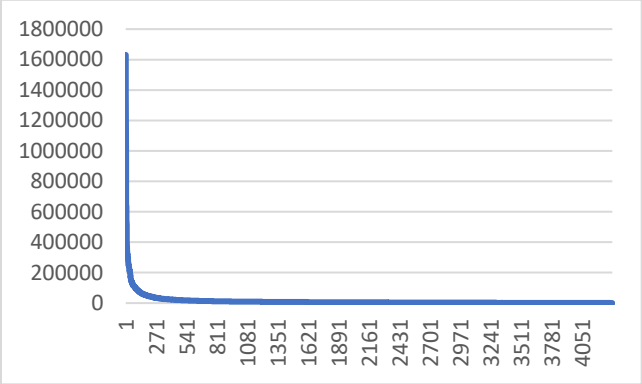

***Figure 2**. Rank (x axis) vs. frequency (y axis) of* Information$_{Dissemination}$ *for each video in the YouTube dataset*

We graph the $I_{Dissemination}$ for all videos in a rank – frequency plot as shown in Figure 2. From Figure 2, the rank – frequency plot follows a classic power law distribution, with a head consisting of a relatively small number of videos that are very popular with many audience segments. Then, there is the long tail consisting of a relatively lot of videos that are popular with a small number of audience segments. This highly skewed pattern of popularity is repeatedly reported in many services (Cha, Kwak, Rodriguez, Ahn, & Moon, 2007).

In fact, when we graph the rank-frequency of $I_{Dissemination}$ in a log-log plot, we find $I_{Dissemination}$ results in the expected power law outcome of a nearly straight line, as shown in Figure 3.

As shown in Figure 3, the log-log plot gives us a nearly straight line, with a slight negative slope, although with some curvature at the ends. A trend line (the dashed line) is a linear equation to which the data points conform quite nicely (R^2 = 0.9574).

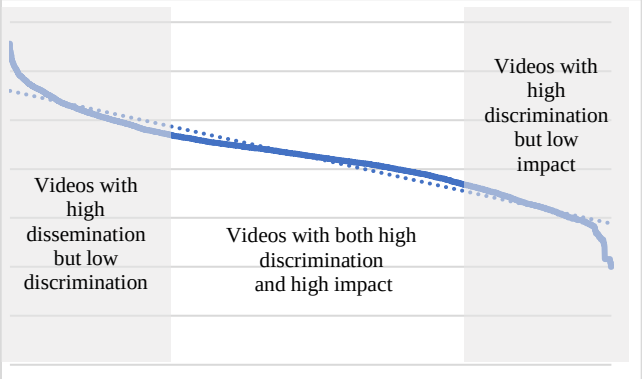

***Figure 3.** Log-log plot of rank (x axis) vs. frequency (y axis) of* Information$_{Dissemination}$ *for each video in the YouTube dataset*

In Figure 3, we have also highlighted and labeled three areas of interest. To the far left, we have the videos with high $I_{Dissemination}$, which are videos with high dissemination and therefore low discriminative power. To the far right, we have the videos with low $I_{Dissemination}$, which are videos with high discriminatory factors but which have low impact, so the audience segments are not meaningful. In our dataset, the volume of views correlates nearly perfectly with a number of audience segments. This investigation of these boundaries leads us to define both τ_L and τ_U.

We first factor out videos with very high $I_{Dissemination}$, which we defined as any video with an $I_{Dissemination}$ equal to or greater than 0.50. This results in 34 videos being removed, representing 25.3 percent of total *views*. We then factor out videos with high $I_{Discrimination}$ but low impact, which we defined as any video with a *view* percentage equal to or less than 0.02 percent of total *views*. This results in 3,649 videos (84.5 percent) being removed, representing 29.7 percent of total *views*.

These results are consistent with expectations, given the power law distribution of our content collection. There are a few videos each with a large frequency of interaction, so their total frequency is quite sizeable. Conversely, there are a lot of videos that individually have few interactions, but collectively represent a sizeable portion of the content but with fewer interactions.

Case Study of the Most Impactful Content

Our analysis to this point leaves us with a set of key content items that have reasonable values for both $I_{Dissemination}$ and $I_{Discrimination}$. Collectively, we identified 637 videos (15 percent of our original dataset) representing 45.0 percent of *views*. These 637 are the key content within the larger collection for discerning meaningful audience segments. As a case study of the applicability of our information dissemination approach for making key business decisions, we focus only on these 637 key videos. Again, we define an audience segment as a unique combination of (*country*, *gender*, *ageGroup*). Using these key videos, we now determine those that are most beneficial, identifying those videos that are particularly impactful in identifying the most unique audience segments.

Our set of 637 videos resonates with users from 170 countries. So, with two gender groups, seven age groups and 170 counties, we have an upper limit of 2,380 audience segments. (i.e., 2 x 7 x 170), assuming that all age groups are represented in each audience segment, which we want to assume in order to identify the videos with discerning impact, So, in this analysis, we use 2,214 as the baseline for maximum audience segments ($Segments_{Total}$).

Unique Audience Segments	Videos	Percentage
0	605	95.4%
1	13	2.1%
2	2	0.3%
3	3	0.5%
4	3	0.5%
5	-	-
6	1	0.2%
7	1	0.2%
8	2	0.3%
9	2	0.3%
10	-	-
12	2	0.3%
Total	637	100.0%

Table 1. *637 of the most impactful content in terms of both a large number of views and number of distinct audience segments*

We then identify the number of unique audience segments, $Segments_{Information}$, for each of our 637 videos, relative to the other videos in the set, shown in Table 1. We find that the most selective audience attribute for videos is the *country*. We identify 18 videos (2.8%) with maximum impact, meaning each audience segment for these videos is unique. However, the *view* for these videos is still somewhat low, which is an outcome of the fractal nature of power law distributions (i.e., nearly every segment of a power law distribution is, in itself, also a power distribution with head and tail.). However, we have 32 videos that specifically and meaningfully identified only one unique audience segment. Naturally, the specific number many change.

Discussion of Data Analysis

Using actual user data from AJ+, an online media outlet, we validate our premise concerning the discriminatory value of information dissemination for determining audience segments for content distributed online. There are several implications for our approach. The major value is the method provides a straightforward yet effective technique for greatly simplifying a complex data set, allowing decision makers to focus on the key content aspects.

In our case, for example, we reduce the dataset by 85 percent, allowing detailed focus and analysis on the ley 15 percent of the dataset that was most impactful in terms of audience segmentation and achievement of a critical KPI.

We were able to determine the most impactful content in terms of resonating with both unique audience segments and individual users. While this is preliminary work, the research results are quite promising with potential impact for understanding how various users interact with online information (at least online videos), aiding content publishers in both producing information that users want to read and, simultaneously, getting content that certain audiences ought to read disseminated in an appealing manner to that audience segment.

Also, we note that the skewed popularity pattern, so-called *long tail*, in AJ+ is not so unique, but is prevalent in various web content services, such as YouTube, Amazon, Netflix, etc. Thus, our findings utilizing such pattern have high generalizability to identify user segments for various services for content publishing.

IMPLEMENTATION OF FINDINGS

We demonstrate the practicality of this line of analysis via use in implementation of a working system (Jung et al., 2017). Our development objective is to develop a system for mining aggregated large scale user data from major social media platforms (e.g., YouTube) in order to identify both distinct and meaningful user segments, and then dynamically generate personas that represent these key user segments. Persona generation is a perfect application for the dissemination analysis presented above, as we need to identify both distinct and impactful audience segments in order to generate the personas.

A persona is a representation of a segment of actual users, presented as an imaginary person. The persona concept is used in IT-development, system design, and marketing. The ultimate artifact is typically a persona description embodying attributes of the user segment that the fictionalized person represents. Personas are an extension of efforts from a variety of domains for identifying, assessing, and constructing groups of people (i.e., users, customers, audience, or market segments) in order to optimize some performance metric (i.e. the speed of task or ease of use). Personas are allegedly well integrated into current design processes (Dharwada, Greenstein, Gramopadhye, & Davis, 2007; Eriksson, Artman, & Swartling, 2013; Friess, 2012; Judge, Matthews, & Whittaker, 2012; Nielsen & Hansen, 2014) for both long and short term projects (Judge et al., 2012).

To dynamically build personas, our methodology requires a multiple-step approach, consisting of:

a) identifying distinct user interaction patterns from the data, using the methodology presented above
b) associating these distinct user interaction patterns to user demographic groups,
c) categorizing meaningful user demographic groups from the data,
d) generating skeletal personas via demographic attributes, and
e) enriching the skeletal personas to create rich personas description.

We initially developed a matrix representing users' interaction with the online products. We denote by V the g×c matrix of g user groups (G_1, G_2, ..., G_g) and c contents (C_1, C_2 ,...,

C_c). The element of the matrix V, V_{ij}, is any value that represents the interaction of user group G_i for content C_j. Using this matrix as the basis, we can identify first distinct user behavior patterns (which can be patterns of any set of user touch points) and then the meaningful user segments from this set of distinct user patterns.

After we have the matrix V, discovering the underlying latent factors, the user interaction patterns that will become the basis of the personas is the next step. There are numerous ways to decompose a given matrix; for this proof-of-concept, we used nonnegative matrix factorization (NMF) (Lee & Seung, 1999) (see Figure 5). Compared to a simple clustering of user groups (Jisun An, Jwak, & Jansen, 2017; J. An, Kwak, & Jansen, 2017), NMF has advantages in that it can find multiple behavioral patterns even from a single group.

The matrix decompositions work as following:

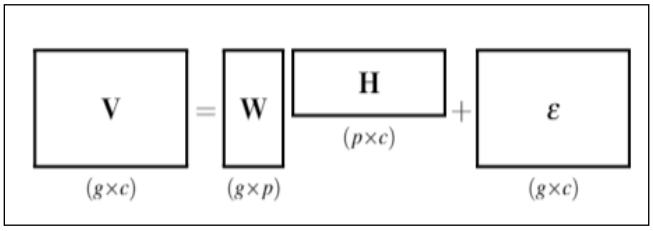

Figure 4. *Matrix decomposition, which identifies distinct interaction patterns and then impactful demographic segments, which form the basis for the resulting personas*

Where **W** is a $g \times p$ matrix, **H** is a $p \times c$ matrix, and ε is an error term. Here, p is the number of latent factors that we can choose. The column in **W** is a basis, and the column in **H** is an encoding that consists of coefficients that combine with each basis and represent a linear combination of the basis. So, W = (demographics x personas). **H** = (personas x contents)

However, there are two factors that must be considered prior to this, which are:

- which meta-hit content to discard, in that content which all demographics viewed is not a discriminator (i.e., nearly everyone watches the certain videos), and
- which demographic groups to discard because they are so small that they are not impactful in any meaningful manner.

In actually generating the personas, we use the method presented above to figure out the content that is too popular (i.e., viewed by too many audience segments) and the content that rarely interacted with (i.e., a very small number of views).

Once done and after application of NMF, we can then generate personas that have distinct sets of user touch points on content and that are also impactful in terms of being meaningful demographics. A set of personas for a major online news organization is shown in Figure 5.

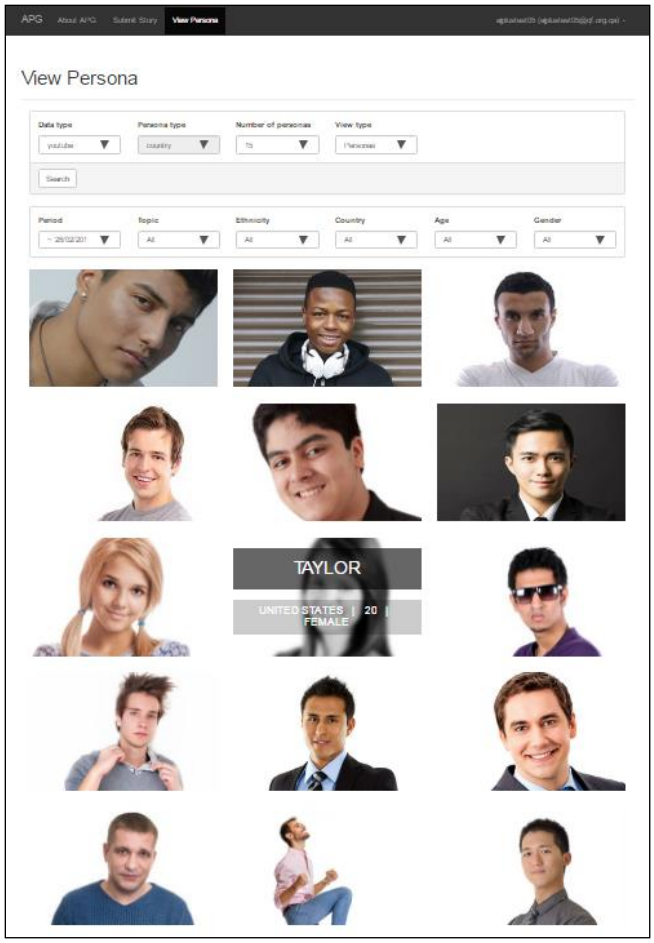

Figure 5. *Set of 15 personas generated based on distinct user touch points and impactful audience demographic size. Mouse over a persona to display persona attributes.*

As shown in Figure 5, a user who wishes to see the audience segments for YouTube, can select the number of personas to generate (the system is currently set for a minimum of five and a maximum of 15). The selected number of personas is then generated. By mousing over one of the personas images, the basic persona attributes are displayed, in this case "Taylor, United States, 20, Female."

The listing of the personas, combined with the mouse overs, afford the user the ability to get an overview of the audience segments in terms of base demographic information. The user can also change content collections, change the number of personas, or apply different filtering, such as data set, topics of interest, ethnicity, country, age or gender.

Each of the images in this persona listing is clickable. Clicking on one icon generates the completed persona description, as shown in Figure 6.

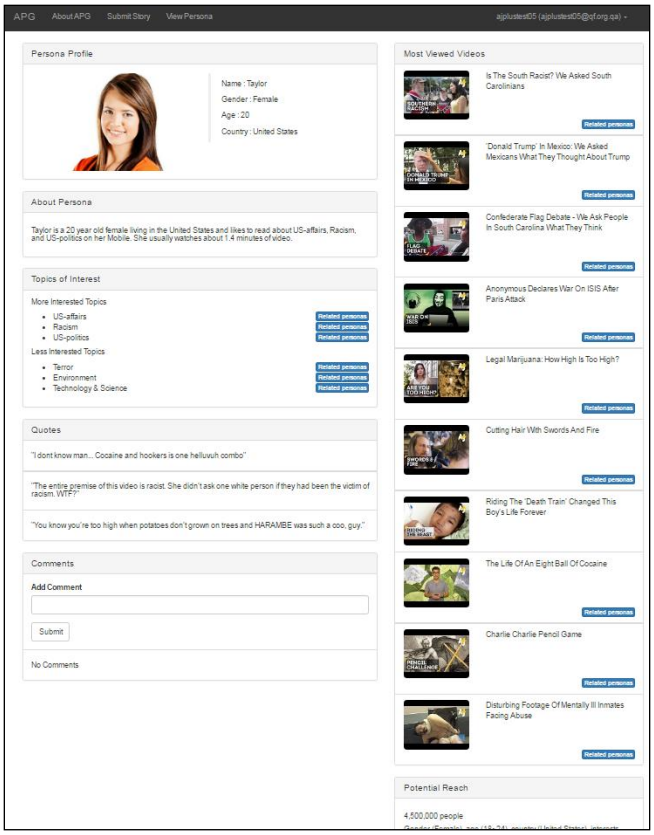

Figure 6. *Persona description with persona attributes, including distinct content viewed*

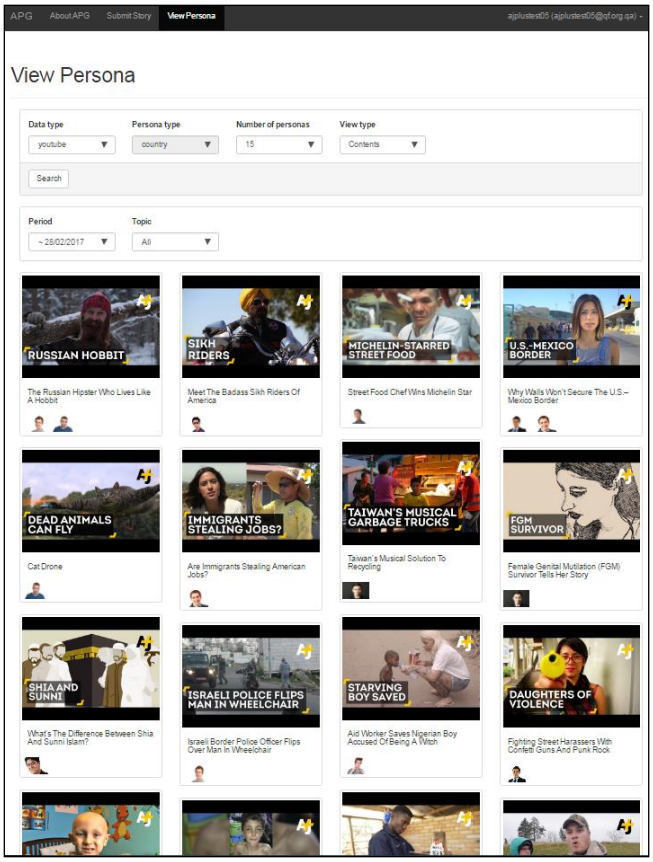

Figure 7. *Content results listing with personas for which content resonates*

DISCUSSION AND IMPLICATIONS

Certainly, there is much additional work to be done, such as a more rigorous definition of determining tau (beyond straightforward heuristics) and the integration of additional KPIs to define the behavioral aspects of users. However, the strength of the research is that we use real content and real audience data from a major social media platform to investigate our concepts for identifying content that can be leveraged to categorize both distinct and impactful audience segments via content dissemination and discrimination.

Our approach, in some sense, is an adoption of the concept of tf-idf for an audience segmentation area. A word is a video, and a document is a demographic group. Following that analogy, term frequency (i.e., video views) maps to the number of views of a video by a given demographic group, and inverse document frequency maps to $Segments_{Information}$. (Thus, $I_{Dissemination}$ because $Segments_{Total}$ is constant).

Then, just as we search a document by a keyword, we can search for a demographic group via a video. The resulting demographic group is a compromise between the high contribution of the view counts and the uniqueness of preference. However, our work is unique in its application of these concepts, of both volume and distinctness, to this context.

CONCLUSION

In this research, we show that the concepts of information dissemination and information discrimination can be used as underlying constructs for meaningful audience segmentation based on content products, when combined with specific performance metrics. Our research shows that audience segmentation can be accomplished rapidly and dynamically using large scale, real time, user data from major online social media platforms, resulting in analysis that reflects the behavior of real people.

Although specifically focusing on digital content, our approach is flexible, resilient, and can be applied in a wide range of contexts. While we limited our focus here to understanding the meaningful audience segments with current levels of information dissemination, it would be interesting to apply the approach to the long tail of audience segments to investigate possible patterns of information diffusion (Rogers, 1976), cultural differences (Salminen et al., 2017), and other social media news aspects (AL-Smadi, Jaradat, AL-Ayyoub, & Jararweh, 2017).

ACKNOWLEDGMENTS

Our thanks to AJ+ of the Al Jazeera Media Network for their wonderful partnership for this research.

REFERENCES

AL-Smadi, M., Jaradat, Z., AL-Ayyoub, M., & Jararweh, Y. (2017). Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic. *Information Processing & Management, 53*(3), 559-750.

An, J., Cho, H. Y., Kwak, H., Hassen, M. Z., & Jansen, B. J. (2016, 29 November-2 December). *Towards Automatic Persona Generation Using SocialMedia.* Paper presented at the The Third International Symposium on SocialNetworks Analysis, Management and Security (SNAMS2016), The 4th International Conference on Future Internet of Things and Cloud, Vienna, Austria.

An, J., Jwak, H., & Jansen, B. J. (2017, 31 July-3 August). *Personas for Content Creators via Decomposed Aggregate Audience Statistics.* Paper presented at the Advances in Social Network Analysis and Mining (ASONAM 2017), Sydney, Australia.

An, J., Kwak, H., & Jansen, B. J. (2017, 4-7 January). *Automatic Generation of Personas Using YouTube Social Media Data.* . Paper presented at the Proceedings of the 50th International Conference on System Sciences (HICSS-50), Waikoloa, Hawaii.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y. Y., & Moon, S. (2007). *I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system.* Paper presented at the Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, Chicago, IL.

Chen, L., Holsapple, C. W., Hsiao, S.-H., Ke, Z., Oh, J.-Y., & Yang, Z. (2017). Knowledge-dissemination channels: Analytics of stature evaluation. *Journal of the Association for Information Science and Technology, 68*(4), 911-930. doi:10.1002/asi.23725

Dharwada, P., Greenstein, J. S., Gramopadhye, A. K., & Davis, S. J. (2007, 1-5 October). *A Case Study on Use of Personas in Design and Development of an Audit Management System.* Paper presented at the Human Factors and Ergonomics Society Annual Meeting Proceedings, Baltimore, Maryland.

Eriksson, E., Artman, H., & Swartling, A. (2013). *The Secret Life of a Persona: When the Personal Becomes Private*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI2013), Paris, France.

Friess, E. (2012). *Personas and Decision Making in the Design Process: An Ethnographic Case Study*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI2012), Austin, Texas, USA.

Fu, W. W., & Sim, C. C. (2011). Aggregate bandwagon effect on online videos' viewership: Value uncertainty, popularity cues, and heuristics. *Journal of the American Society for Information Science and Technology, 62*(12), 2382-2395. doi:10.1002/asi.21641

Hendahewa, C., & Shah, C. (2017). Evaluating user search trails in exploratory search tasks. *Information Processing & Management, 53*(4), 905-922.

Judge, T., Matthews, T., & Whittaker, S. (2012). *Comparing Collaboration and Individual Personas for the Design and Evaluation of Collaboration Software*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI2012), Austin, Texas, USA.

Jung, S., An, J., Kwak, H., Ahmad, M., Nielsen, L., & Jansen, B. J. (2017, 6-11 May). *Persona Generation from Aggregated Social Media Data.* Paper presented at the ACM Conference on Human Factors in Computing Systems 2017 Extended Abstracts (CHI2017), Denver, CO, USA.

Lee, D. D., & Seung, S. H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788-791.

Mellor, V. (2006). *Mastering audience segmentation: How to apply segmentation techniques to improve internal communication* London: Melcrum.

Nielsen, L., & Hansen, K. S. (2014). *Personas is Applicable: A Study on the Use of Personas in Denmark*. Paper presented at the Proceedings of the 32nd Annual ACM conference on Human factors in Computing Systems (CHI2014), Toronto, Ontario, Canada.

Ortiz-Cordova, A., & Jansen, B. J. (2012). Classifying web search queries in order to identify high revenue generating customers. *Journal of the American Society for Information Sciences and Technology, 63*(7), 1426-1441.

Rogers, E. M. (1976). New product adoption and diffusion. *Journal of Consumer Research, 2*(1), 290-301.

Salminen, J., Sengün, S., Kwak, H., Jansen, B. J., An, J., Jung, S.-G., . . . Harrell, F. (2017, 21-23 Aug.). *Generating Cultural Personas From Social Data: A Perspective of Middle Eastern Users.* Paper presented at the The Fourth International Symposium on Social Networks Analysis, Management and Security (SNAMS-2017). Prague, Czech Republic.

Smith, W. R. (1956). A product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing, 21*(1), 3-8.

Song, X., Lin, C.-Y., Tseng, B. L., & Sun, M.-T. (2005). *Modeling and predicting personal information dissemination behavior.* Paper presented at the Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05), New York, NY.

Stern, B. B. (1994). A revised communication model for advertising: multiple dimensions of the source, the message, and the recipient. *Journal of Advertising, 23*(2), 5-15.

Tan, L. K.-W., Na, J.-C., & Ding, Y. (2015). Influence diffusion detection using the influence style (INFUSE) model. *Journal of the Association for Information Science and Technology, 66*(8), 1717-1733. doi:10.1002/asi.23287

Wang, C., Zhou, Z., Jin, X.-L., Fang, Y., & Lee, M. K. O. (2017). The influence of affective cues on positive emotion in predicting instant information sharing on microblogs: Gender as a moderator. *Information Processing & Management, 53*(3), 721-734.

Zhang, L., Zheng, L., & Tai-QuanPeng. (2017). Structurally embedded news consumption on mobile news applications. *Information Processing & Management, 53*(5), 1242-1253.