Automatic Generation of Personas Using YouTube Social Media Data

Jisun An
Qatar Computing Research Institute, HBKU
jan@qf.org.qa

Haewoon Kwak
Qatar Computing Research Institute, HBKU
hkwak@qf.org.qa

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

Abstract

We develop and implement an approach for automating generating personas in real time using actual YouTube social media data from a global media corporation. From the organization's YouTube channel, we gather demographic data, customer interactions, and topical interests, leveraging more than 188,000 subscriber profiles and more than 30 million user interactions. We demonstrate that online user data can be used to develop personas in real time using social media analytics. To get diverse aspects of personas, we collect user data from other social media channels as well and match them with the corresponding user data to generate richer personas. Our results provide corporate insights into competitive marketing, topical interests, and preferred product features for the users of the online news medium. Research implications are that reasonably rich personas can be generated in real time, instead of being the result of a laborious and time-consuming manual development process.

1. Introduction

The use of personas, employed in marketing and advertising for abstract user representation, has spread to a variety of other fields, such as system design and content creation. A persona is a representation of a group or segment of users that share common behavioral characteristics. A persona is developed in the form of an explicit but fictitious individual representing a segment of consumers, users, or stakeholders. The fictitious person is accompanied by a detailed narrative that represents a set of consumers possessing similar behaviors, attributes, or characteristics.

To be beneficial, a persona must be perceived as real. Therefore, in order to make the fictitious individual appear as a real person to the content producers or product developers, the persona representation can contain a variety of demographic data, behavioral details, socio-economic status, gender, along with other richness details such as hobbies, family members, friends, or possessions. Likewise, the description of the persona can address the goals, needs, wants, frustrations, and other emotional aspects of the fictitious individual that are pertinent to the content or product being designed. Also, the persona is usually given a name and an image to assist the content developers in focusing on the appropriate user segment [1].

Although there is increasingly online data concerning consumers and market segments for an organization, personas are still mainly developed via ethnography methods, such as focus groups or surveys, which have inherent issues concerning freshness, reliability, representativeness, cost, and timeliness. In this work, we demonstrate that one can generate reasonable rich and continually updated personas in real time using online social media and other data by incorporating social media analytics.

2. Related work

Personas are beneficial for guiding development decisions and evaluating project ideas, from systems to services to advertisements. Personas have long been used for assisting marketing decisions and content creation in advertising. Within the domain of system design, personas are considered to be a useful technique in directing designers and developers.

Typically, a persona is created for clarifying and synthesizing descriptions of user segments, with the idea that a persona assists in focusing on the behavior patterns, wants, and needs of a particular segment of users. Actual products that are deployed to market can have multiple user segments that require various personas, so the development of the granularity of data and execution of data integration in a meaningful way can often be complex.

Ideally, one wants the construction of a persona based on actual data and research of a product's intended user base or market demographics. However, this has been and is a major issue in persona development, as the user data has commonly been gathered via ethnography methodologies, such as

URI: http://hdl.handle.net/10125/41250
ISBN: 978-0-9981331-0-2

surveys or focus groups [2]. Common problem with creating personas via these methods is that the personas are many times not based on first-hand user data [3], the data set is not of a sample size that can be considered statistically significant, and the personas cannot be easily updated. Also, utilizing these processes for persona development can be costly and time-consuming [4], resulting in personas that are slow to take to market.

As a result of these shortcomings, in practice, persona development does not typically come from current user behavioral data. Instead, personas are based on the assumptions, experiences, or expectations of executives, marketers, consultants, or designers. Even if the personas are developed from actual user data, the personas can rapidly become out of date, as the user data is not continually updated via reoccurring surveys or focus groups. Therefore, organizations rely on personas that are not believable or do not actually represent the genuine current or targeted users [5]. These problems are exacerbated in actual commercial products in fast moving and competitive market areas, where the user base is multivariate and open to the possibility of flux. Automated persona generation can address these issues.

3. Research objective

We believe that social media analytics that analyzes patterns in social media data to enable informed decision-making [7] can provide competitive advantage to organizations via the automatic generation of personas. In order to achieve such an objective, it means that personas (1) must represent the current users of the product and (2) be sensitive that usage can dramatically change based on audience interests or shifts over time [8]. In this work, we propose that actual consumer behavior and related demographic data concerning users of a product, service, or content [9][10] can be rapidly and inexpensively collected from a variety of social platforms and analyzed in order to generate personas in real time. Earlier work in this direction is reported in [23][24][25].

4. Data collection

We validate our research premise using actual user data from AJ+, a global media corporation. Perhaps surprisingly, audience preferences have been somewhat ignored by journalists in the news industry mainly because of the lack of accurate measurements. This issue was especially true before the era of online news, but it continues today. Several studies point out large differences between news production and consumption patterns using, for example, the number of views of articles in news websites [11]. The correct understanding of audiences becomes more important to increase marketshare and consumption of digital content, which online audience data can provide [12]. Being an online digital content producer, AJ+ also creates actual commercial products (i.e., news articles, in particular, short news video clip) in a marketplace with intense competition, complex consumer demographics, and potentially rapidly changing customer trends. Prior work has shown that news article topic, article content, media platform, and individual factors together drive consumers engagement [13].

Therefore, we consider AJ+ an excellent organization for both data collection and analysis. Our research with actual AJ+ user data shows the value of automatic persona generation for news readership research in particular, although we consider the results transferable to other industry verticals where the understanding of audience preferences is key for success.

4.1. Data collection organization: AJ+

Our data collection organization, AJ+ (http://ajplus.net/) is an online news channel from Al Jazeera Media Network that is natively digital with a presence only on social media platforms, including a mobile application for major smartphone devices. The media concept is unique in that AJ+ was designed from the ground up to serve news in the medium of a viewer, versus a teaser in one medium redirecting to a website via a hyperlink. Given that AJ+ is based on social platforms with a presence on iOS and Android apps, the digital content is specifically designed to be viewed in the Facebook newsfeed, YouTube Channel, or Twitter Timeline for the audience segments that are most active on those platforms.

Also, AJ+ has been very innovative in its experimentation with storytelling formats, app design, and video development, receiving significant press [14]. At the time of this study, AJ+ was the second largest producer of video on Facebook, had more than 3 million Facebook followers, 195,000 Twitter followers, and 188,000 YouTube subscribers. Also, in a testimony to their digital content production, AJ+'s engagement rate at the time of the study was 600 percent (i.e., their news products are engaged by 6x their follower base), which is a fantastic reach for digital content products.

Therefore, given the prerequisite for rapid and media specific development of content in a competitive and fluid information market [6][15], AJ+ has a critical

need for automatic and real time generation of personas to guide digital content, media, and system planning and design, along with tracking consumer trends.

Consequently, in pursuit of our overall research objective of automatically generating personas in real time, we are specifically interested in understanding the AJ+ audience by identifying (1) whom are they reaching (i.e., market segment) and (2) what content are associated with each market segment, focusing here on the digital content. From these research results, combined with other user data, we can generate personas in real time using actual user data, and keep the personas updated and current.

4.2. YouTube Data Collection

For data collection, we primarily focus on the YouTube channel in the research reported here, although we do include Twitter and Facebook features in the system development. The main reason to focus on this site is that the YouTube analytics platform gives the most detailed statistics for every video, compared to other channels.

The YouTube analytics platform provides, for each of AJ+ videos, user profile data (e.g., gender, age, country location, and which site the user comes from) at an aggregate level. We leverage these datasets to include demographic information into our automatic persona generation system. We access the data in real time, meaning that the data is continually collected.

The statistics in the YouTube analytics platform can be accessed by YouTube APIs[1]. Although there are other metrics besides viewCount (the number of views) and viewerPercentage, such as likes or comments, the sheer value of the number of likes or comments is much lower than that of views. Thus, we focus only on viewCount and viewerPercentage. We note that the data is private and available only to an owner of the YouTube channel (i.e., AJ+), and thus not publicly accessible.

4.3. Exploratory Data Analysis

We first present some overall statistics from the AJ+ YouTube channel page. At the time of the study, there were 2807 AJ+ videos posted. These videos had more than 30 million views by users from 217 countries.

The AJ+ audience is worldwide, with the top three countries in terms of viewership being Canada, Great Britain, and the United States (US), each representing 2.44 percent of total viewership in terms of the number of unique videos watched. In terms of the total number of views, the US is the largest country market segment, with about 49.4 percent of views coming from the US. It is interesting to note that, although AJ+ was designed to target the US market, a majority of viewers come from outside the US.

In terms of gender and age distribution (13-17 years, 18-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, and 65 years and older), 20.9 percent of viewers were female, with 79.1 percent being male. Regarding the age, as expected, young adult male is the biggest segment.

Some videos showed a worldwide appeal, with 100 videos being viewed in 100 or more countries. Conversely, there were about 100 videos that were viewed by users from 5 or fewer countries. In terms of the actual number of views, the viewership counts per individual videos follow a power law distribution, with a small number of videos being viewed a lot and a large number of videos being viewed a small number of times. Such skewed popularity of videos is one of the well-known characteristics in YouTube [16].

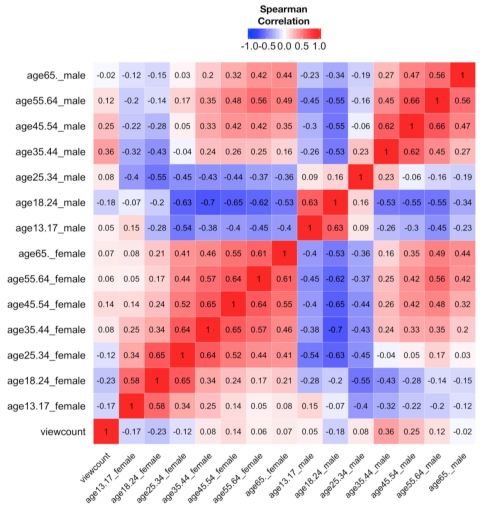

Figure 1 Spearman rank correlation between topics and demographic attributes

We then examine which demographic groups are similar to each other in terms of their video consumption patterns. For each of demographic groups, we rank all videos by the viewCount of the group. Then, we compare this ranked list to that of other groups by the Spearman ranking correlation coefficient. We plot the result in Figure 1.

Findings show that users that are of similar age are likely to consume similar content, overall. Within the ``Female'' group, as the age difference increases, the correlation decreases. For the ``Male'' group, there is a clear distinction among age groups. The Age 13-17

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

male group has a high correlation with Age1 8-24 male group (0.63), meaning the popular videos in one group is also popular in the other group. However, Age 13-17 male group has very low correlation with Age 25-34 male group (< 0.09), indicating their consumption patterns are very different. Interestingly, older males have somewhat similar tastes with females in the similar age grouping, while younger males show very different patterns with females in all ages.

Given the distribution and diversity of the dataset, the results from the exploratory data analysis indicate that AJ+ could benefit from personas representing various key customer segments.

5. Methodology for Automated Persona Generation

For persona generation, we first cluster user groups with similar behavioral patterns. Then, to make the persona perceived as a real person in the system, we extract demographic data and add a variety of additional information to the persona, such as a name, a photo, topical preferences, top videos consumed, and quotes.

5.1. User Clustering

In order to automatically generate personas from the YouTube social media data, our approach was first to identify sets of users who share common behavioral characteristics. Then, we identify distinct demographic characteristics of those user clusters to create the personas.

The YouTube analytics platform does not provide statistics on the video consumption of a specific individual but that of a certain user group, determined by the combination of (gender, age, country). For example, it shows that the user group of (Female, 13-17, Korea) accounts for 307 views for video V_1, 204 views for video V_2, etc. The possible values of (gender, age, country) are:

- Gender: male / female
- Age: 13-17 years, 18-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, and 65 years and older
- Country: one of 198 possible counties

As a result, the number of possible demographic groups of [Gender, Age, Country] is 2,772, and we cluster these groups based on their news video consumption patterns. It is worth mention again that we use the entire audience of AJ+, thus reflecting the skewed gender ratio. As our goal is to correctly understand the whole audience, we do not sample users to consider gender balance. Instead, when we cluster demographic groups, we use video viewing patterns only. Thus, the unbalanced gender distribution does not affect the clustering process.

To get more general content consumption patterns, we transform video-level viewing patterns into topic-level ones. For such transformation, we classified the topic of each AJ+ video by using its title and short description (normally a few sentences). We used the Alchemy Taxonomy API for topic classification[2]. For a given text, the Alchemy Taxonomy API returns suggested topics and confidence scores. We take the topic with the highest confidence score. In particular, Alchemy Taxonomy API supports over 1,000 topics, and they are well-organized as a hierarchical structure.

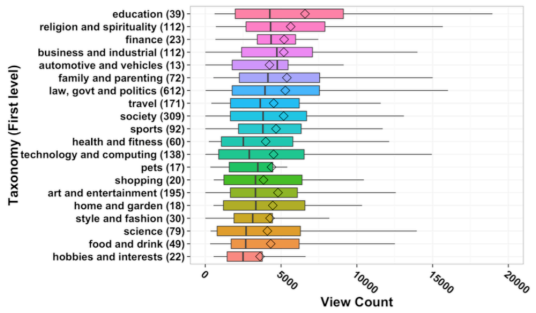

Figure 2. Viewcounts of AJ+ videos according to their topics

Based on our analysis, AJ+ videos are classified into 22 tier-1 level topics. Among them, we remove two topics (career and real estate), with less than 10 videos each. The viewCount for each resulting topic is shown in Figure 2. The top three most covered topics by AJ+ are `law, government and politics' (612), `society' (309), and `art and entertainment' (195). In contrast, the top three most viewed topics are `education', `religion and spirituality', and `finance'.

For clustering, we use a hierarchical clustering method. The hierarchical clustering method builds a dendrogram among entities and clusters entities based on the dendrogram. Its main advantage is that the dendrogram is computed only once regardless of the number of clusters of interest. Thus, it is appropriate for the situation where the number of clusters can be varied by user input and the clusters need to be found on demand. This is the exact setting we assume for the automatic persona generation system. We build a G×T matrix where G is the number of demographic groups and T is the number of topics. We calculate a group-topic weight $W_{\{gi,tj\}}$ as follows:

1) Select a set of videos categorized as a given topic, t_j.

[2] http://www.alchemyapi.com/products/alchemylanguage/taxonomy

2) Compute the proportion of the group g_i's viewCount of the videos (instead of a raw viewCount).
3) Choose a median of those values as a group-topic weight, $W_{\{gi,tj\}}$.

We use this method to avoid a possible bias in clustering output due to the skewed distribution of the popularity of videos. The issue is, as we explained in the previous section, that the popularity of the AJ+ content on YouTube follow a power-law distribution, meaning that some videos get viewed a lot and others, relatively, not much at all. In this case, if we use the raw viewCount for clustering, it results in one giant cluster including many groups and a few small clusters. There could be other ways to resolve this issue, such as removing top 5% and bottom 5% videos ranked by viewCount. However, regardless, the raw viewCount for any video can be an important feature to characterize groups. Taking this aspect into account, we design our approach to treat all videos equally important. This, together with topic-level aggregation, resolves the problems raised by the skewed behavior data.

In building the matrix, we consider the top 50 countries ranked by the number of total viewCounts across all videos, and we have 20 Alchemy's tier-1 level topics. This results in 50×20 matrix with matching group-topic weight values. We apply hierarchical clustering method on this resulting matrix, building a hierarchy of clusters. Once the dendrogram is built, we can simply choose the number of clusters (k) of our interest.

Here, we describe, as an example, the resulting ten clusters k equals 10) by the topic preference. We quantify the level of interests of a cluster (c_k) to a topic (t_p), denoted as a cluster-topic weight $W_{\{ck,tp\}}$). Note that a certain set of demographic groups are now related into one of the ten clusters. Then, we compute the cluster-topic weight by taking the average of the group-topic weights ($W_{\{gi,tj\}}$ where $g_i \in c_k$).

For comparison, we standardize a set of cluster-topic weight values by each cluster, $W_{\{ck,\,T\}}$ where T is a set of all topics (20 in our case). We compute Z-score based on the mean (μ) and the standard deviation (σ) of $W_{\{ck,\,T\}}$ as follow:

$$Z = \frac{W_{\{c_k,T\}} - \mu}{\sigma}$$

We note that a positive Z-score means the value is higher than the mean μ) of the cluster-topic weight values, and a negative Z-score means the value is lower than the mean. We then use the standardized cluster-topic weights for the comparison of topic preference across clusters. From here, we refer to these clusters as the personas, as we give them other contextual information and features.

Figure 3 shows the bar charts of the standardized cluster-topic weights for each topic for three example personas. Persona 1 shows particular interests in two topics: 'hobbies and interests' and 'family and parenting'. Persona 2 is likely to read more about 'shopping'. The results indicate that the resulting personas well reflect groups of similar topical interests.

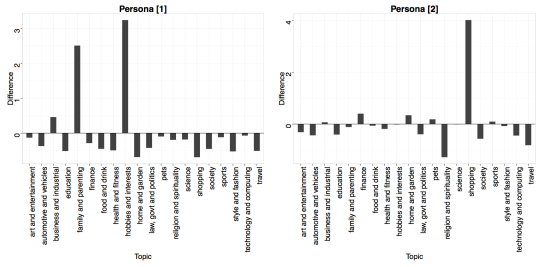

Figure 3. Topic preference by persona

To understand how demographic segments are clustered within each persona, we use a network science approach. For each country, the number of demographic segments is 2×7=14, which is a combination of gender (2) and age groups (7). We build a network whose nodes are these 14 segments, and edges are whether two demographic segments are in the same persona. We then superimpose N networks built from N countries. In the resulting network, the weight of an edge is the number of countries that have the corresponding edge. As expected, since the resulting network is the complete (fully connected) network, we need to focus on more important edges. Thus, we use backbone extraction proposed by [17]. We get one big connected component and two isolated nodes (male age 65- and female age 55-64). We plot the connected component in Figure 4.

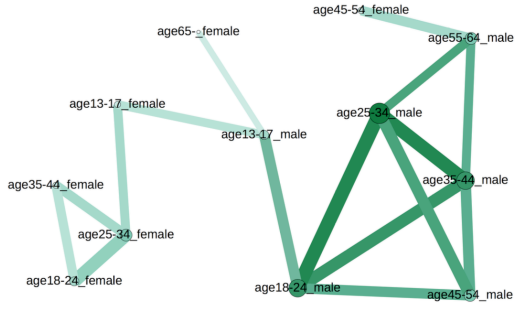

Figure 4. Relationship between demographic segments in Personas (k=10)

We find two categories that consist of three (Age 18-24 Female, Age 25-34 Female, and Age 35-44 Female) and four (Age 18-24 Male, Age 25-34 Male, Age 35-44 Male, and Age-45-54 Male) nodes, respectively. For both categories, the youngest

generation (13-17) and the older generation are not included in the groups. In other words, young adulthood and middle age group share their taste of video news consumption, while minors and older generation have different taste. This tendency did not clearly emerge the in Spearman correlation analysis previously shown in Figure 1, where we observed some age groups of males and females have a quite similar taste of video consumption. Here, we find the contrasting but more important pattern of gender segregation through network science. This finding implies that understanding of gender difference, along with topical preferences, in content consumption is key to generating accurate personas.

5.2. Name and photo matching

Once we have the results of clustering method, we now attach additional information, such as a name and a photo, to each persona, adding richness so that it can turn to the standard representation of a persona. To do this, we build a dictionary of names and photos for each segment of (gender, age, country). There are existing dictionaries, which are usually maintained by the government of each country, of popular names for gender and the birth year (e.g. https://www.ssa.gov/oact/babynames/). We collect them to build our dictionary for matching age appropriate names with specific personas.

Similarly, we purchased the copyrights to high-quality photos for each of all demographic segment (gender, age group, ethnicity) of interest and develop a photo library for persona generation.

We then, from each cluster of users, find a representative demographic group that has the highest number of aggregated viewCounts in the cluster. Then, we match names and photos for that particular demographic group, such as (male, 18-24, US) or (female, 24-35, UK).

For example, if the representative demographic group in the cluster is a 55-year-old US female, the name for the fictitious person is one of common names of female babies who are born in 1961 (2016-55=1961), US. Similarly, we present a photo matching the representative demographic group in both gender and ethnicity.

5.3. Preference

We also derive a list of preferred videos and less preferred videos for each user persona. As the videos can be traced into topics based on the video descriptions, we have a list of preferred topics and less preferred topics for each persona. This is a good starting point of the general preference of the fictional character represented by the persona. We can get more information by combining other social media user data, which we introduce in more detail in the following sections.

5.4. Quotes Generation

A quote is a short phrase that introduces oneself or expresses one's opinion and is widely used as one trait of a persona. To generate quotes for the persona, we bring additional user information from other social media channels, specifically Twitter and Facebook, and match them with corresponding personas demographics, leveraging the richness of user data in social media. Here we collect publicly available information only. For privacy concerns, we do not use sensitive information, such as actual user name or photo. Thus, we believe that we minimize the risk of privacy issues.

To match users across different social media, we demographically characterize users in each social media channel and then find users who have similar characteristics.

We first estimate the demographic information of Twitter users. Here, we collect a list of Twitter users who live in a certain country by using Followerwonk[3]. Followerwonk indexes self-reported location fields in user profiles on Twitter and enables searching for Twitter users by location. Then, we infer the gender of the users by Face++[4]. Face++ analyzes profile photos and infers the age and gender by convolution neural network and for gender, it achieves more than 90% accuracy [18]. Lastly, for age, we look at the profile description (i.e., bio) of Twitter users. We look at users who explicitly mention their age (e.g., "26 yr|yrs|year|years old"). Then, we compile a list of descriptions for the combinations of (country, age, gender) and match them with the corresponding persona that we generated from the YouTube data.

5.4. Other Traits

Combining multiple social media user data has much potential to make a persona richly nuanced. We demonstrate how we extract quotes from Twitter, but what we can obtain by combining multiple social media data can be varied. For instance, Kosinski et al. demonstrate that the Likes data can be used to predict a wide range of personal attributes, such as religion, political views, happiness, parental separation, etc. by studying over 58,000 Facebook users [19]. We are

[3] https://moz.com/followerwonk/
[4] http://www.faceplusplus.com/

investigating extracting such information from various social media channels.

In contrast to Twitter, collecting user information from Facebook is tricky because of the privacy issue. Most Facebook users do not make much, to any, personal information publicly available. However, one possibility to extract demographic information from Facebook is using publicly shared links. For example, using the description of each of shared links from Facebook users, we can detect user languages and topics. Likewise, these shared links could reveal user's economic status. Prior research has shown that affluent customers (high-end luxury product websites) and budget conscious customers (price aggregation or discount websites) can be distinguished by websites they visited [20]. We are currently investigating extracting rich information beyond user topical interests from shared Facebook links.

6. Validation

We demonstrate the value of the automatically generated personas in two ways: (a) first we validate the automatically generated personas by comparing them to personas generated via ethnography methods by AJ+ themselves, and (b) we then validate the stability of our personas by using the consistency of the clusters.

First, we compare the automatically generated personas with personas generated via ethnography methods done by AJ+ themselves. By interviewing AJ+ producers and editors, we have heard how they shape their digital content products using their personas. When AJ+ was founded, a consultant was commissioned to identify the new channel's primary audience and generate personas for the newly formed news agency. The consultant proposed three US-based personas: Kathleen (Female, 34), Deigo (Male, 28), and Kelly (Female, 32).

From our analysis, of actual AJ+ consumer data, the gender-age demographic of these three personas, in total, represent only 16.60 percent of the AJ+ viewing US audience on the YouTube platform. Worldwide, these personas represent only 8.20 percent of the AJ+ viewing audience. Obviously, producers and editors relying only on these personas for digital content production would be reaching only a small segment of their actual customer base, would be misdirecting products, and be ignoring sizable market segments. These comparisons also highlight the major issues of traditionally generated personas, in that one is unsure if they actually represent the true customer base at a given time. We also attempted to compare topical interests, but the demographics were so far off that it was not fruitful.

Second, we show the stability of generated personas by examining the consistency of the clusters. We split our dataset into 80% (dataset I) and 20% (dataset F). We then cluster demographic groups based on viewing patterns for each of the datasets I and F separately, and we compare them to see how consistent are the clusters.

We denote $D=\{D_1,D_2,\ldots,D_N\}$ as a set of demographic groups. For two different demographic groups, D_i and D_j ($i \neq j$), we can define $S(D_i,D_j)$, which is a pairwise metric to check whether two demographic groups are in the same cluster, as following:

$$S(D_i, D_j) = \begin{cases} 1 & \text{if } D_i \text{ and } D_j \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

Superscript of $S(D_i,D_j)$ indicates which dataset is used for the computation. For example, S^I(20s,30s) is 1 if 20s and 30s are falling into the same community in the dataset I.

Then, we can define the consistency for any pair of demographic group D_i and D_j between the dataset I and F.

$$C^{I \leftrightarrow F}(D_i, D_j) = \{S^I(D_i, D_j)\} \text{ XNOR } \{S^F(D_i, D_j)\}$$

, where XNOR is the logical complement of the exclusive OR gate (e.g. 1 XNOR 1 = 1, 0 XNOR 0 = 1, and otherwise 0). In other words, this measure checks whether a pair of demographic groups that are consistently found in the same cluster or consistently found in different clusters.

Finally, we summarize $C^{I \leftrightarrow F}$ for every D_i and D_j as follow:

$$C^{I \leftrightarrow F} = \frac{\sum_{i=1}^{N} \sum_{j=i+1}^{N} C^{I \leftrightarrow F}(D_i, D_j)}{|\{(D_i, D_j) | 1 \leq \forall i, j \leq N \text{ and } i < j\}|}$$

, where the denominator is for the normalization. The range of $C^{I \leftrightarrow F}$ is between 0 and 1. When the clustered partitions are identical between I and F, the value of $C^{I \leftrightarrow F}$ becomes 1. The condition when the minimum value is observed is rather complicated; when the number of clusters in I is 1 and that of clusters in F is N (or vice versa) is the only condition when $C^{I \leftrightarrow F}$ is zero.

For the different number of clusters from 2 to 20, we calculate $C^{I \leftrightarrow F}$. To better understand how good or bad the consistency is, we standardize $C^{I \leftrightarrow F}$. We create 1,000 random ensembles, R_i, of clusters and compute Z-score based on the mean (μ) and the standard deviation (σ) of $C^{I \leftrightarrow R_{1..1000}}$ as following:

$$Z = \frac{C^{I \leftrightarrow F} - \mu}{\sigma}$$

We note that cluster size distribution in any R_i is the same as the actual distribution of cluster size in F.

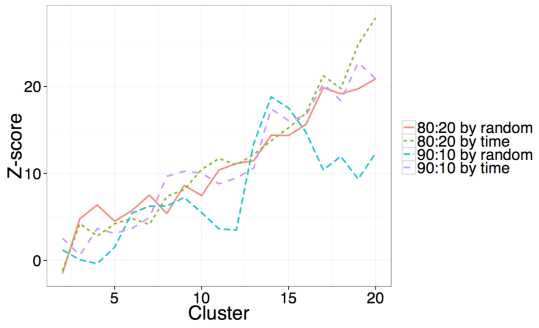

Figure 5. Standardized score of $C^{I<->F}$ with different number of clusters

Figure 5 shows how the standardized score, Z, changes under the different number of clusters. We also test the different division proportion (90% and 10%) and the different division method (by random and by time). It is clear that our clustering is much more consistent than random assignments because the Z-score is quite high. Also, interestingly, we find that no division scheme shows consistently higher or lower Z-score with varying number of clusters. We find 14 out of 19 cases (k=2 to 20) that the temporal 90% and 10% split has a higher Z-score than the random 90% and 10% split, and 11 out of 19 from the 80% and 20% split. This shows the importance of the persona generation with the more recent data.

The high consistency of generated personas firmly proves the stability of our approach, although it also indicates that the AJ+ audience did not vary significantly during the period of data collection. We further discuss the usability and usefulness of our approach in the discussion section.

7. System building

In this section, we present and discuss the current state of system development. Our research aim is to use actual user behavior, rapidly collected and analyzed, to generate personas in real time. So, we have developed a robust system that automatically generates personas representing current consumers of AJ+ news content based on the methods we explained above. The system updates these personas in real time based on any changes in audience demographic, interests, or usage.

In our system, one can select the attributes that one desires to generate personas around, currently gender-age-country, and various levels of granularity. For news organizations, one can also filter the personas by topics, which for AJ+ is one of the twenty news article classifications. By selecting all attributes and the maximum number of personas, the system generates the results with an example shown in Figure 6.

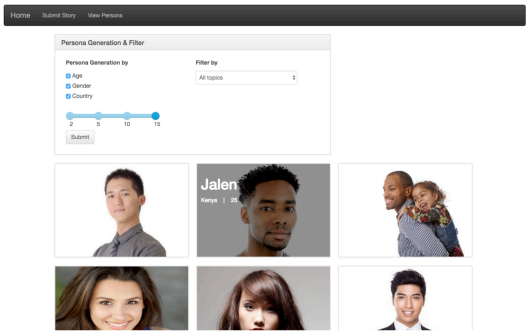

Figure 6. Screenshot of automated persona generation system for resulting personas

From Figure 6, multiple personas are presented, with names, ages, and pictures assigned in real time. The pictures, with all copyrights purchased, are gender, age, and ethnically appropriate. The names for the personas are also age appropriate. A mouse rollover causes the persona's name, age, and country to appear (see Figure 6, Jalen persona).

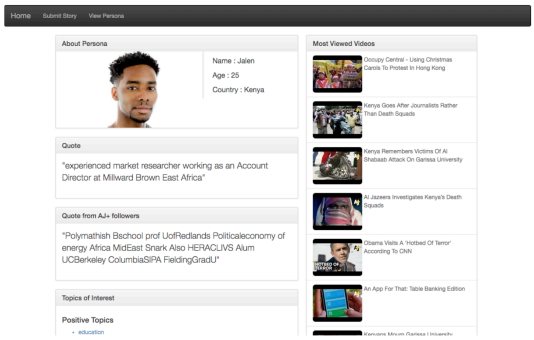

Figure 7. Screenshot of a single persona

As shown in Figure 7, the persona heading appears (e.g., Jalen). There are currently two quotes, one from the corresponding over Twitter population (e.g., Quote) and one from the specific AJ+ Twitter populations (e.g., Quote from AJ+ followers). Specifically, for AJ+'s news video focus, the top and bottom Topics of Interests are displayed. Finally, the top ten most watched videos by this persona are also presented.

8. Discussion and implications

We have demonstrated several important findings and outcomes. Most importantly, the current automatic persona generation system is quite robust, pulling in millions of social media interactions from YouTube and other social media showing that one can use actual online user behavior and demographic data that is rapidly collected and analyzed in order to generate personas in near real time. Nevertheless, there are several research and development fronts that we are pursuing.

Overall, the biggest strength of our approach is that we benefit from actual user data, both demographic and behavioral, reducing time and cost for generating personas relative to traditional methods, and thus our approach is suitable for real time persona generation. Also, with our approach, we do not need to carefully sample users for interviews to develop personas. Instead, we can extract representative personas from millions of users by using online social media, combined with other data.

8.1. Scalability of the system

Scalability of the persona generation system may be the possible concern of our approach, as there could be hundreds of thousands to millions of records. Anticipating this, we have carefully designed the current system architecture to scale up and handle this level of user data. Currently, our system cumulatively accesses and stores the YouTube analytics data using the YouTube API.

8.2. Benefit for journalists

Concerning benefits to organizations, our persona systems currently provides demographic information, rich behavioral details, and specific customer interests. Our system design approach is also derived from collaboration with journalists who actually benefit from the generated personas as the journalists create new digital content. The journalists require realistic views of the actual users so that they can reach those targeted readers with better titles, content, and article framing. Our persona research offers a strong foundation for achieving this objective.

We validate the stability of our approach by confirming the consistency of clusters. At the same time, this high consistency of the generated personas implies the stable audience of AJ+. It might decrease the value of the real timeliness, one virtue of the data-driven approach, of our automatic persona generation because the generated personas of AJ+ have not changed a lot. However, the real timeliness would be worthwhile in particular scenarios. Consider the situation where one AJ+ video or article becomes viral and a new audience comes to the AJ+ channel. Producers need to quickly understand who they are and what they are interested in, besides the viral item, to keep their attention. Our approach helps to handle such situations by quickly generating personas only of new comers in real time.

9. Conclusion

In this work, we have taken the initial fruitful steps to move persona creation from a manual, time-intensive, staleable process to one that is automated, in real time, and current. Presently, we integrate user data primarily from one social media platform, supplemented with data from other services, analyze this data, and automatically generate personas beneficial to, in this case, journalists. For future research, we are continuing system development in order to enhance features for personas selection, persona filtering, and multiple topics. Additionally, we want to do a more indepth evaluation of the system with actual journalists. Moreover, we are also investigating leveraging other data sources to provide richer demographic attributes, attitudinal characters, and other aspects for rounding out the generated personas. We would like to extend personas attributes to social interactions [21] among cluster members. We fully believe that leveraging social media and other online data can create business success [22]. Given the millions of users interacting with social media channels on social media platforms, we believe that this data can cost-effectively generate personas to target current and potential consumers.

10. References

[1] T. Matthews, T. Judge, and S. Whittaker, “How do designers and user experience professionals actually perceive and use personas?” in Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2012, pp. 1219– 1228.

[2] J. Pruittand, T. Adlin, The persona lifecycle: keeping people in mind throughout product design. Morgan Kaufmann, 2010.

[3] N. Khalayli, S. Nyhus, K. Hamnes, and T. Terum, “Persona based rapid usability kick-off,” in Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2007, p. 1771 1776.

[4] J. McGinn and N. Kotamraju, “Data-driven persona development,” in Proceedings of the SIGCHI conference on

human factors in computing systems. ACM, 2008, p. 15211524.

[5] T. Judge, T. Matthews, and S. Whittaker, "Data-driven persona development," in Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2012, pp. 1997–2000.

[6] S. Faily and I. Flechais, "Persona cases: a technique for grounding personas," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2011, pp. 2267–2270.

[7] N. Bekmamedova, and G. Shanks, "Social media analytics and business value: A theoretical framework and case study," in Proceedings of the 47th Hawaii International Conference on System Science. ACM, 2014, pp. 3728–3737.

[8] M. Goul, S. Balkan, and D. Dolk, "Predictive analytics-driven campaign management support systems," in Proceedings of the 48th Hawaii International Conference on System Sciences. ACM, 2015, pp. 4782–4791.

[9] B.J. Jansen, K. Sobel, and G. Cooks, "Classifying ecommerce information sharing behaviour by youths on social networking sites," in Journal of Information Science. ACM, 2011, pp. 120–136.

[10] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth networking sites," in Journal of Information Science. ASIST, 2009, pp. 2169– 2188.

[11] S. Abbar, J. An, H. Kwak, Y. Messaoui, and J. Borge-Holthoefer, "Consumers and suppliers: Attention asymmetries. a case study of al jazeeras news coverage and comments," in Proceedings of the Computation+Journalism Symposium, 2015.

[12] G. Shuradze and H.-T. Wagner, "Towards a conceptualization of data analytics capabilities," in Proceedings of the 49th Hawaii International Conference on System Sciences. ACM, 2015, pp. 5051–5063.

[13] E. Mao and J. Zhang, "What drives consumers to click on social media ads? the roles of content, media, and individual factors," in Proceedings of the 48th Hawaii International Conference on System Sciences. ACM, 2015, pp. 3405– 3413.

[14] J. Roettgers, "How al jazeeras aj+ became one of the biggest video publishers on facebook," in Variety. 5051–5063. ACM, 2015, pp.

[15] R. Sinha, "Persona development for information rich domains," in Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2003, pp. 830–831.

[16] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007, pp. 1–14.

[17] M. A´. Serrano, M. Boguna´, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," Proceedings of the national academy of sciences, vol. 106, no. 16, pp. 6483–6488, 2009.

[18] H. Fan, M. Yang, Z. Cao, Y. Jiang, and Q. Yin, "Learning compact face representation: Packing a face into an int32," in Proceedings of the ACM International Conference on Multimedia. ACM, 2014, pp. 933–936.

[19] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," Proceedings of the National Academy of Sciences, vol. 110, no. 15, pp. 5802–5805, 2013.

[20] A. Odlyzko, "Privacy, economics, and price discrimination on the internet," in Proceedings of the international conference on Electronic commerce. ACM, 2003, pp. 355–366.

[21] J. Wang, C. Chiang, C. Chang, and P. Chung, "Groupsona: A method for discovering the needs of social interaction embedded services," in Proceedings of the 44th Hawaii International Conference on System Sciences. ACM, 2011, pp. 1530–1605.

[22] C. K. Coursaris, W. V. Osch, and B. A. Balogh, "Do facebook likes lead to shares or sales? exploring the empirical links between social media content, brand equity, purchase intention, and engagement," in Proceedings of the 49th Hawaii International Conference on System Sciences. ACM, 2016, pp. 3545–3554.

[23] J. An, H. Y. Cho, H. Kwak, M. Z. Hassen, and B. J. Jansen, "Towards automatic persona generation using social media," in The Third International Symposium on Social Networks Analysis, Management and Security (SNAMS 2016), The 4th International Conference on Future Internet of Things and Cloud. IEEE, 2016, p. Workshop Paper 2.

[24] J. An, H. Kwak, and B. J. Jansen, "Validating social media data for automatic persona generation," in The Second International Workshop on Online Social Networks Technologies(OSNT-2016), 13th ACS/IEEE International Conference on Computer Systems and Applications AICCSA 2016. IEEE, 2016.

[25] B. J. Jansen, J. An, H. Kwak, M. Z. Hassen, and H. Y. Cho, "Efforts towards automatically generating personas in real-time using actual user data," in Poster presented at Qatar Foundation Annual Research Conference 2016. QF, 2016, p. ICTPP3230.