2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Personas for Content Creators via Decomposed Aggregate Audience Statistics

Jisun An
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
Email: jisun.an@acm.org

Haewoon Kwak
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
Email: haewoon@acm.org

Bernard J. Jansen
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
Email: bjansen@hbku.edu.qa

***Abstract*—We propose a novel method for generating personas based on online user data for the increasingly common situation of content creators distributing products via online platforms. We use non-negative matrix factorization to identify user segments and develop personas by adding personality such as names and photos. Our approach can develop accurate personas representing real groups of people using online user data, versus relying on manually gathered data.**

I. Introduction

The persona, introduced to the design domain by [1] with follow-up refinement [2], is a representation of an actual set of users presented as an imaginary person describing user goals. Personas are a continuation of efforts from a variety of domains for identifying and assessing groups of people to optimize some performance metric (e.g., speed of task or ease of use). Personas are well integrated into current design processes [3]. The recommended number of personas per project historically has been no more than a handful.

A known problem is that creating personas is not a cheap or quick procedure, as the creation has historically involved ethnography methods. As one time data collection events, the personas created can be quickly outdated. Without real time data, designers have no confirmation whether the personas are representative of current users.

While there have been some online data-driven approaches [4] [5], they have used fine-grained user-level data that is not often available and potentially having privacy issues. As such, approaches using individual level data are not suited for most content creators, who see aggregated statistics via a social media platform's analytic tools.

Our research focus is to automatically create personas using of real user data that has been *aggregated* by social media platforms [6]. The domain of digital content creation has the unique but increasingly common attribute of having access to aggregated user data, rather than session level data, which complicates the persona generation process.

These limitations are the motivations for our research, which we propose, develop, implement, and evaluate an approach for mining privacy-preserving aggregated statistics of user interaction with digital content and then enriching the results with personal attributes in order to generate rich personas. The biggest contribution of this approach is that it can be easily adapted for the actual content creators, since such aggregated user statistics are the de facto standard provided by social media analytic tools.

ASONAM '17, July 31 - August 03, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07$15.00
http://dx.doi.org/10.1145/3110025.3110072

II. Framework to Develop Personas

Our approach begins with one matrix encoding users' interactions to content. We denote by $\mathbf{V}$ the $g \times c$ matrix of g user groups ($G_1, \ldots, G_g$) and c contents ($C_1, \ldots, C_c$). The element of the matrix $\mathbf{V}$, V_{ij}, is any statistic that represents the interactions of user group G_i for content C_j.

Our approach is generalizable and applicable across 1) data of different granularity and 2) any content type. In matrix $\mathbf{V}$, a user group, G, can be an individual user if the data is available and a privacy concern does not exist, meaning our approach is applicable to both user-level and aggregated-level data. In addition, many social media analytic tools provide data that can be easily turn into the matrix $\mathbf{V}$. For example, YouTube Analytics and Facebook Insights offer statistics of interactions (e.g. view counts) by demographic groups for each video and post. Thus, any content provider using those platforms can easily use our approach regardless of the type of content (e.g., apps, articles, videos, or images).

A. Non-Negative Matrix Factorization to Identify Behavioral Attributes for Personas

Once we have the matrix $\mathbf{V}$, next is to discover the underlying latent factors, the content consumption patterns, which become the "basis" of the personas, by the matrix decomposition as following:

$$\underset{(g\times c)}{\mathbf{V}} = \underset{(g\times p)}{\mathbf{W}} \underset{(p\times c)}{\mathbf{H}} + \underset{(g\times c)}{\boldsymbol{\varepsilon}} \quad (1)$$

where $\mathbf{W}$ is a $g \times p$ matrix , $\mathbf{H}$ is a $p \times c$ matrix, and ε is an error term. Here, p is the number of latent factors (behavioral pattern or skeleton personas) that we can choose. The column in $\mathbf{W}$ is a basis, and the column in $\mathbf{H}$ is an encoding that consists of coefficients that combine with each basis and represent a linear combination of the bases.

NMF does not allow negative entries in $\mathbf{W}$ and $\mathbf{H}$. A column in $\mathbf{H}$ represents each of common content consumption patterns. A coefficient, H_{ij}, shows the importance of content, C_j, to explain the content consumption pattern, P_i. A row in $\mathbf{W}$ represents each user group consisting of different common consumption patterns. The coefficient, W_{ij}, is a relative proportion of a consumption pattern, P_j, in a user group, G_i. The result of NMF becomes a set of "skeleton" personas, which we turn into rich personas by adding personality.

B. Adding Personality

1) Persona End Goal: As mentioned, $\mathbf{H}$ shows a set of common content consumption patterns represented by a linear combination of contents. Then, we can infer the end goal of the users with a specific consumption pattern using domain knowledge. For example, if a person watches videos of new games for PlayStation 4 but not for XBox One, then, we can assume an end goal of this person is to buy new games for PlayStation 4.

2) Persona Demographics: One important property of the persona is demographic characteristics. For that, we take a two-step approach: (a) finding a representative user group for a "skeleton" persona and (b) identifying the representative demographics of this group. First, we focus on $\mathbf{W}$ in equation (1). A row in $\mathbf{W}$ represents how each user group can be characterized by different consumption patterns. A column in $\mathbf{W}$ shows how a common consumption pattern is associated with different user groups. Thus, for each column, the user group with the largest coefficient can be interpreted as a "representative" user group for that corresponding pattern. Next, we determine the demographics of the representative user groups. This depends on how one defines user groups in $\mathbf{V}$. The most efficient way is to use the data broken-down by demographics when building $\mathbf{V}$. If a row in $\mathbf{V}$ maps to a group defined as (age group, gender, country), a format which social analytic tools often provide, then it is trivial to find representative demographics of a persona.

3) Persona Name: We build a dictionary of names by collecting popular baby names by gender and year from many countries. For example, the top 1,000 popular baby names any year since 1879 (https://www.ssa.gov/oact/babynames/) is available for US. Then, through age group, gender, and location of a representative group, we can assign a temporal, gender, and ethnically appropriate name to a persona.

4) Persona Photo: To assign a photo to a persona, we purchase copyrights to commercial photos of models for different ethnicities, genders, and ages. Here, the selection of different styles of figures to represent different professions, interests, etc. can strengthen the expressive power of the persona, so we selected varied photos for each and tag each photo with the appropriate meta-data. Then, through age group, gender, ethnicity, etc. of a representative segment, we can assign an appropriate photo to a persona.

5) Other Personas Details: Beyond these persona description properties, we can leverage online user data to add additional personal details. The core idea is that, for a given persona, it is possible to (i) find a set of *similar* users in other social media, and (ii) extract and *blend* their personal details from their publicly available contents (e.g. tweets) in order to address privacy considerations.

III. Application: Identifying Target Persona of Content Prior to Publication

One of the benefits of using NMF for generating personas is a clear association, represented in $\mathbf{H}$ ($p \times c$), between personas and interests, or non-interest, in digital content. Beginning with this association, we can identify target personas of new content, $\mathbf{H}_{\mathrm{n}}$, even before content publication.

The problem of predicting interest in new content has been studied in recommendation systems. The most intuitive solution is to find "similar" content relative to the new content and approximate that the level of interest in similar content will remain the same. To compute the similarity of content in a robust way, we define content features. The features can be anything: topics, length, price, and so on. Formally, we define a matrix, $\mathbf{F}$ (c contents $\times$ f features), capturing the features of content. We then can derive another matrix, $\mathbf{K}$ (p personas $\times$ f features), represents an association between a persona and content features:

$$\mathbf{K} = k(\mathbf{H}, \mathbf{F}) = \varphi(\mathbf{H})\varphi(\mathbf{F}) \tag{2}$$

where k is a kernel function and some appropriate mapping function φ. For computational simplicity, here we assume φ=$\mathbf{I}$. In other words, the interest in content is the sum of the interest in its features. Then, we can get a simple multiplication of two matrices:

$$\mathbf{K} = \mathbf{HF} \tag{3}$$

By multiplying $\mathbf{F}^{-1}_{\mathrm{right}}$ both sides, we get:

$$\mathbf{H} = \mathbf{KF}^{-1}_{\mathrm{right}} \tag{4}$$

where $\mathbf{FF}^{-1}_{\mathrm{right}}$=$\mathbf{I}$.

The representation of $\mathbf{H}$ in equation (4) guides us how to predict $\mathbf{H}_{\mathrm{n}}$. For new content, we can define $\mathbf{F}_{\mathrm{n}}$ that represents new content and their features. By substitute $\mathbf{F}_{\mathrm{n}}$ into equation (4), we can get $\mathbf{H}_{\mathrm{n}}$:

$$\mathbf{H}_{\mathrm{n}} = \mathbf{K}(\mathbf{F}_{\mathrm{n}})^{-1}_{\mathrm{right}} \tag{5}$$

$(\mathbf{F}_{\mathrm{n}})^{-1}_{\mathrm{right}}$ can be computed by the following:

$$(\mathbf{F}_{\mathrm{n}})^{-1}_{\mathrm{right}} = \mathbf{F}^{\mathrm{T}}_{\mathrm{n}}(\mathbf{F}_{\mathrm{n}}\mathbf{F}^{\mathrm{T}}_{\mathrm{n}})^{-1} \tag{6}$$

when $\mathbf{F}_{\mathrm{n}}$ has linearly independent rows ($\mathbf{F}_{\mathrm{n}}\mathbf{F}^{\mathrm{T}}_{\mathrm{n}}$ is invertible). If not, we split a set of new content into several sets so that $\mathbf{F}_{\mathrm{n}}$ of each set has linearly independent rows. This procedure

avoids the loss of generality of the method. Then, we write a new equation (7) based on (5) and (6):

$$\mathbf{H}_n = \mathbf{K}\mathbf{F}_n^T(\mathbf{F}_n\mathbf{F}_n^T)^{-1} \quad (7)$$

The key of equation (7) is that $\mathbf{K}$, the matrix representing an association between personas and features, does not need to be changed for newer content because $\mathbf{K}$ depends on content features not the content itself. Also, by combining equation (1) and (7), for new content, we can get $\mathbf{V}_n$:

$$\mathbf{V}_n \approx \mathbf{W}\mathbf{H}_n = \mathbf{W}\mathbf{K}\mathbf{F}_n^T(\mathbf{F}_n\mathbf{F}_n^T)^{-1} \quad (8)$$

Through this equation (8), it is possible to predict the views of new content by persona based on content features.

IV. Case Study: Generating Personas for AJ+

We present our experience in working with AJ+, an online news channel from Al Jazeera Media Network. We first collect the audience data from the AJ+ YouTube channel. Using the data offered by YouTube channel analytics, we build $\mathbf{V}$, whose entry, V_{ij}, is a view count of a content (video) by a group that is defined as a combination of age, gender, and country. In our dataset, we have 181 unique countries, two gender groups, and seven age groups, resulting in 181 $\times$ 2 $\times$ 7 = 2,534 groups. We excluded groups with less than 1,000 views in total as they do not reveal meaningful pattern. This results in 281 groups, covering 95% of total views.

Consulting journalists at AJ+ regarding the cognitive limitations in remembering all the details of personas, we choose ten as the best number of personas for AJ+. We highlight that our methodology can be easily rerun for a different number of personas, which makes our framework different than conventional manual persona generation methods.

A. Ten Personas for AJ+ Video Consumers

By applying NMF to $\mathbf{V}$, we can have skeleton personas. Then, we build the rich persona by adding personality.

1) Adding Persona End Goals: We reveal the end goals of each persona from content consumption patterns, in particular, from a set of discriminative videos consumed. The discriminative videos of a persona are defined as the videos for which that persona has a higher probability to watch than another personas would. We identify the discriminative videos for each persona by Chi-square test [7] on $\mathbf{H}$ (p<.05).

We examine the topic of the discriminative videos of each persona to get insight into her purpose in engaging with AJ+ content. To infer the topic of a video, we first attempted to use Alchemy Taxonomy API, but we found that it does not work well mainly due to the conciseness of AJ+ videos titles. Instead, by consulting with AJ+ journalists and using an open coding method, we build a dictionary of keywords for 14 topics and classify AJ+ videos based on the keywords. Some topics are shown in Table I.

We get the fraction of videos of the topic t for a persona i (denoted as $F_t^{(i)}$). Then, using this fraction value, we examine which video topics a persona has a higher tendency to watch compared to the other personas. We quantify such topical preferences by computing the Z-score for $F_t^{(i)}$. For the topic t and the persona i, the Z-score can be computed as $Z_t^{(i)} = \frac{F_t^{(i)} - avg(F_t)}{\sigma}$, where F_t is a set of $F_t^{(X)}$ for any existing persona X, and σ is the standard deviation of the F_t. The higher Z-score means that a persona is more likely to watch videos of the topic than the other personas normally do.

TABLE I
Topics and keywords used for classifying AJ+ videos

Topic	Selected Keywords
Entertainment	nfl, sport, cooking, restaurant, cat, film, messi
Environment	climate, whale, tornado, wildfire, recycle
Refugees	refugee, syria, weiwei, overboard
Tech. & Sci.	alpha go, robot, wheelchair, wikipedia, vr
US-politics	obama, trump, clinton, sanders, gop, pacs

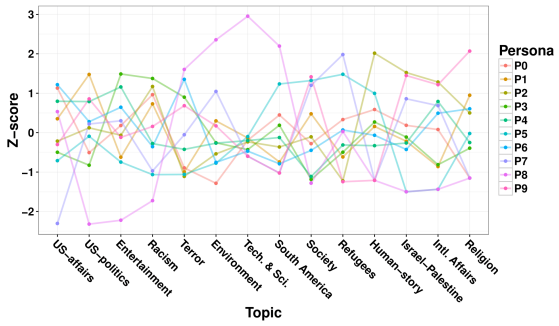


Fig. 1. Frequency of 14 main topics

Figure 1 shows the Z-scores of 14 topics across 10 personas. We note that all ten personas have different patterns, indicating unique topical preferences. As one example, Persona 3 (P3) has high positive Z-scores for Entertainment, Racism, Terror, and South America. In our content analysis, we observed that videos classified as Terror often mentioned the attacks by ISIS in Europe and videos classified as Racism focused on reporting the protests triggered by Ferguson incident. We thus infer and summarize the end goals of this persona as *following the world news and seeking new information regarding recent terror attacks and protests against racism*. Similarly, we infer common end goals (e.g., set of digital content topics) of each of our ten personas.

2) Resulting Personas: The summary of our ten personas along with their names, demographics, personality, topical preferences of AJ+ videos, and their end goals is shown in Table II. Due to lack of space, we show the only five personas and omit photos. We find that three out of ten personas have (Male, U.S.) as a representative group. This makes sense given that more than 75% of the traffic comes from U.S. However, we also note that these three groups are in three different age ranges, and their topical preferences are distinctive. Our ten personas are based on the actual audience statistics of AJ+ in YouTube and thus are well capture its current audience.

Ideally, each persona should be described in a narrative way.

TABLE II
EXAMPLES OF PERSONAS REPRESENTING AJ+ AUDIENCE. SECONDARY INTERESTS ARE SHOWN IN THE (PARENTHESIS).

ID	Name	Gender	Age	Country	Bio	End Goal
P0	Robert	Male	27	U.S.	Op-Ed columnist for several online news magazines.	Mainly checking US politics and affairs. Following refugee and human right issues around the world. (South America and Israel-Palestine)
P2	Benjamin	Male	35	Portugal	Worked in Dubai for years. Has close Muslim friends	Following human rights and refugee stories. Coming for AJ+'s casually for topics from art to tech. (Israel-Palestine and Religion)
P3	Ahmad	Male	25	Malaysia	Night club bomb terror happened in neighboring city.	Following world news and seeking new info. of recent terror attacks as well as protests against racism. (Entertainment and Human-stories)
P6	Sarah	Female	33	U.S.	Mom of a lovely girl. Toward a safe world for children. Freelance Designer.	Concerned about violence against minorities. Following women issues and disasters around the world. Enjoying casual items and human-stories for leisure. (Religion)
P8	Manuel	Male	31	Mexico	Chemistry undergrads in U.S. Part of Mexican start-up	Following current issues in U.S., checking AJ+ for daily news in Texas and Central America, enjoying tech-news. (Environment and Terror)

We have developed a system to automatically generate persona descriptions, presented in [6].

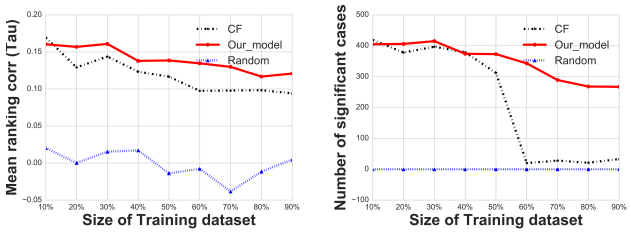
B. View Prediction

Quantitatively evaluating our method, we predict the view counts of demographic groups (a combination of (age, gender, country)) for new videos by using equation (8). We divide all 4,323 videos, ordered by published time, into 10 slices. Among the 10 slices, we use the 10th slice as our testing set, which is the latest 432 videos. Then, for training, we use some of the remaining slices to consider recency and their expressive power. Although there is a general belief that more training data leads to better prediction performance in machine learning, in our case, more videos to train the model is not necessarily helpful because the audience of AJ+ might change over time. In such cases, more older data would not reflect the behavioral patterns of the current audience.

We run experiments with 9 different sizes of the training data sets by changing N from 10% to 90% with offset of 10%. $N = 10\%$ means the ninth slice only, and $N = 30\%$ means the 7th, 8th, and 9th slices. For each training set, we construct a matrix $\mathbf{V}$, and by applying NMF, we get a matrix $\mathbf{W}$ and $\mathbf{H}$. Then, we build a matrix $\mathbf{F}$ and $\mathbf{F}_\text{n}$ by a Latent Dirichlet Allocation (LDA) with 100 topics. With these five matrices, we estimate the view counts of new videos for demographic groups, $\mathbf{V}_\text{n}$ (g groups $\times$ n videos), by equation (8).

For each of new videos, we rank demographic groups based on weight values in $\mathbf{V}_\text{n}$. We compare this ranking with the true ranking of groups by Kendall rank correlation coefficient. We compare our model with two other models: 1) random model and 2) collaborative filtering (CF) model. The random model ranks groups randomly for a new video. The CF model computes the average view counts of each demographic group and uses them for any new video, as CF-based recommending system assigns an average behavior of users for "new" content.

Figure 2 shows: (a) the average τ among 432 test videos where the result is statistically significant ($p < 0.05$) and (b) the number of significant cases. The poor performance of the random model (Figure 2(a)) and having few significant cases (Figure 2(b)) proves that the view counts from each demographic groups for videos is far from the random construction. In Figure 2(a), our model outperforms the CF-based model except when N is 10%. This demonstrates that our persona-based prediction performs well, even though our current implementation has used some approximation like equation (3) with a limited features set.



(a) (b)

Fig. 2. The result of predicting the ranking of demographic groups by Random, CF, and our model.

The performance of the CF-based model drops when $N >$ 40%. AJ+ faced a sudden change of their viewership in that period. The CF-based model is not robust for such sudden changes, while our persona-based approach is robust enough to adapt to the changes of the audience.

V. CONCLUSION

Our research shows that personas can be rapidly and automatically created from large scale, real time, aggregated user data from major online social media platforms, resulting in personas that are based on real data that reflects real people. We evaluate our personas, showing that they have predictive ability. Our approach is flexible for a range of contexts.

REFERENCES

[1] A. Cooper, *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity*. Sams Indianapolis, 1999, vol. 261.
[2] J. Pruitt and T. Adlin, *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufman, 2006.
[3] L. Nielsen and K. S. Hansen, "Personas is applicable: A study on the use of personas in denmark," in *Proceedings of CHI'14*. ACM, 2014, pp. 1817–1823.
[4] X. Zhang, H.-F. Brown, and A. Shankar, "Data-driven personas: Constructing archetypal users with clickstreams and user telemetry," in *Proceedings of CHI'16*. ACM, 2016, pp. 5350–5359.
[5] M.-F. Chiang, E.-P. Lim, and J.-W. Low, "On mining lifestyles from user trip data," in *Proceedings of ASONAM '15*. ACM, 2015, pp. 145–152.
[6] S.-G. Jung, J. An, H. Kwak, M. Ahmad, L. Nielsen, and B. J. Jansen, "Persona generation from aggregated social media data," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2017, pp. 1748–1755.
[7] G. Casella and R. L. Berger, *Statistical Inference*. Duxbury Pacific Grove, CA, 2002, vol. 2.