Creating Manageable Persona Sets from Large User Populations

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

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

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

ABSTRACT

Creating personas from actual online user information is an advantage of the data-driven persona approach. However, modern online systems often provide big data from millions of users that display vastly different behaviors, resulting in possibly thousands of personas representing the entire user population. We present a technique for reducing the number of personas to a smaller number that efficiently represents the complete user population, while being more manageable for end users of personas. We first isolate the key user behaviors and demographical attributes, creating *thin personas*, and we then apply an algorithmic cost function to collapse the set to the minimum needed to represent the whole population. We evaluate our approach on 26 million user records of a major international airline, isolating 1593 personas. Applying our approach, we collapse this number to 493, a 69% decrease in the number of personas. Our research findings have implications for organizations that have a large user population and desire to employ personas.

CHI'19 Extended Abstracts, May 4-9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5971-9/19/05.
DOI: https://doi.org/10.1145/3290607. 3313006

KEYWORDS

Personas; Big Data; Web analytics; user segmentation

Table 1: Key Constructs and Definitions for Data-Driven Personas.

Persona - Fictional person created from data to represent a user type that uses, or might use a site, brand, or product
Persona Description - Portrayal of the persona, typically in 1-2 pages that synthesizes the actual user data
Demographics - Statistical data relating to a population or particular groups within it
User Segment - Groups of individuals that are similar in specific ways, such as age, gender, interests or behaviors
Persona Set - One or more personas needed to represent each segment in the population
Thin Persona - Incomplete persona with demographic (age, gender, country) and behavioral attributes, but missing name and picture.

1 INTRODUCTION

Personas have traditionally been developed using qualitative methods using data collection methods such as surveys and focus groups [11]. Deriving personas from such approaches have typically resulted in a handful of personas [10]. However, with the increased availability of online user and customer data, known as Big Data [15], there is both the opportunity and need to use data-driven personas derived directly from a system's users or a company's customer analytics data. This data-driven persona creation is a major shift from the way personas have traditionally been developed [6], and it offers benefits as either a stand-alone method or in conjunction with other tactics, such as creating hybrid personas from both quantitative and qualitative data.

A related outcome from the availability of a large amount of customer and user data is the *increase in the number of personas*. With major systems and businesses being online, user and customer populations in the millions are not uncommon. As such, it is often unrealistic to expect a handful of personas, which have traditionally been used [1], to adequately represent user populations of such magnitude. Additionally, data-driven persona development allows the direct linking from personas to individual user data, i.e., there is a persona that represents each user available in the data. Such an increase in the size of the user population combined with the way personas have traditionally been crafted using a somewhat standard set of demographic attributes, such as gender, age, nationality, and behavior creates the issue that organizations can end up with a large number of personas. Obviously, this number is often too high for both effective and efficient deployment by human decision makers.

In this research, we address this issue via a two-level algorithmic approach that:

a) identifies the unique user behavioral patterns using a matrix decomposition approach,
b) develops *thin personas* by linking these behavioral patterns with demographical attributes,
c) collapses this set of *thin personas* using a clustering cost function,
d) produces the minimal number of *thin personas*, and
e) generates the complete persona profiles for the reduced set.

We find that we can reduce a very large set of personas into a more manageable set using the above algorithmic approach. In our data, as a proof of concept, we achieve a 69% reduction in the number of personas reducing the set of personas from 1,593 to 493 while still representing every user segment in the population with a persona. Although 493 is still a large amount, it is reasonable given the size of the user population. To further narrow down the number of personas, we suggest giving the end users of personas filtering options, so that they can find relevant personas on-demand for their use case. See Table 1 for key variables and definitions.

2 REVIEW OF LITERATURE

However, there has been no research that we could locate investigating what is the appropriate number of personas, although a set of 3 to 6 seems a de facto guideline. In general, it seems that each user segment should have a representative persona. Segmentation is the conceptual foundation

Table 2: Behavioral and Demographic Variables in the Dataset.

Key Behavioral Flight Variables
Top Ten Flight Destinations (Destination 01, Destination, 02, … Destination 10
Behavioral Variables
Booking Reference Code
Booking Date
Booking Channel used in creating the bookings (Group, Online, Others)
Point of Sale City (where the flight was purchased)
Flight Number
Flight Date
Destination City
Cabin Class (F - First, J - Business, Y - Economy)
Demographic Variables
Age
Gender
Nationality

Behavioral variables, that we used in this study, are presented in Table 2. Along with these behavioral variables are demographical variables (see Table 2).

All personally identifying values were masked in the dataset, not available to the researchers.

of data-driven personas that employ actual user data [8, 13]. Segmentation is defined as the sub-division of a population into homogeneous sub-sections of individuals [3, 8], where any sub-each user segment should have a representative persona. Segmentation is the conceptual foundation of data-driven personas that employ actual user data [8, 13]. Segmentation is defined as the sub-division of a population into homogeneous sub-sections of individuals [3, 8], where any sub-section is distinct from other segments [4]. However, what is a homogenous segment is somewhat debatable, as there are many attributes that can define a user segment [5], including behavioral and demographic ones. A typical set of demographic attributes, from online social platforms, is gender, age, and country [2]. Behavioral attributes can be a wide assortment of variables, depending on the domain and focus [9, 14]. With data-driven personas derived for major online systems, this combination of the standard demographic attributes and domain-specific behavioral attributes can result in thousands of personas, which is far more than can be employed effectively in organizations in the manner personas are typically used.

3 RESEARCH OBJECTIVES

Our research objective is: *Develop a technique to reduce a set of personas to the minimum number required to represent all segments in the user population.*

For this research, we define the minimum number of personas as 'the smallest number of personas where each persona is behaviorally and demographically distinct from any other persona and where the set of personas represents all segments in the user population'. Behaviorally unique is a set of one or more behaviors that frequently occur together, forming an exclusive group of behaviors. In prior work [12], we have outlined how to isolate unique user behavioral patterns from large complex sets of user data and how to link these unique user behavioral patterns to user demographical attributes in order to create *thin personas*. A thin persona is a combination of demographic and behavioral attributes but no individualized information such as a photo or name typically seen in a persona profile.

For each set of unique behaviors, there can be one or more demographic groups that exhibit these behaviors. Each demographic group is defined by a set of demographic variables, which on many of the major social media platforms include gender, age group, and country, but can also include others, such as nationality and city. However, some of these demographic groups associated with a set of unique behaviors might not be distinct, in that they can be collapsed with adjoining demographic groups associated with same behavior set. We determine the demographic distinction using a cost function that measures the dissimilarity with adjoining groups, as explaining in the following section.

4 METHODOLOGY

Our dataset has more than 26 million records from a major international airline company of flight bookings over a two-year period (see Tables 2 and 3). We use this dataset to demonstrate how our persona collapsing methodology reduces the data-

Table 3: Locked and Collapsible Variables Used to Determine Number of Personas.

Locked Variables
Booking Reference Code
Booking Date
Booking Channel
Flight Number
Flight Date
Destination City
Cabin Class
Collapsible Variables
Age (13-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65+)
Gender (Male, Female)

Although one could use many approaches for determining which variables to collapse, we identified these two variables as being collapsible via factor analysis, showing that Age was not a good predictor of behavior and Gender was a weak predictor. Nationality was a fairly strong predictor of behavior, so we kept it as a *locked variable*.

Table 4: Definitions of Key Concepts.

Conceptual definitions
Thin persona – persona profile with basic attributes
Persona - persona profile with full set of attributes
Collapsible variable – variables with values can be merged with adjacent values
Locked variable – variables with values cannot be merged
Interaction matrix -matrix where passengers are rows, flight destinations are columns, and cells are the number of bookings)

base, and we use it to demonstrate how our persona collapsing methodology reduces the number of personas while still covering the behavioral archetypes within the data.

4.1 Identification of Unique Behavioral Patterns

As our approach for identifying unique behavioral patterns is explained thoroughly in prior work [12], we only briefly present it here. We rely on non-negative matrix factorization (NMF), which is a well-known type of matrix decomposition. Our approach begins with an interaction matrix encoding user behaviors, defined here as booking a flight from a point of origin to a destination. Once we have the interaction matrix, we discover the number of latent interaction patterns by decomposing it; explaining the persona's preference toward a set of destinations. We set the number of destinations within the set to ten for each pattern for ease of presentation. For identifying the demographic customer segments, we find the set of representative demographic segments for each behavioral segment. The end product is a set of the unique behavioral patterns each associated with one of the sets of user demographics, defined as (Age, Gender, Nationality). For this research, use the top ten most frequent behavioral patterns, representing nearly 96 percent of the user database.

4.2 Identification of Extraneous Personas

After the generation of *thin personas* (see Table 4), we employ an algorithmic approach to reduce the number of personas to the minimal necessary that still represents the user population. To accomplish the collapsing, we employ a cost function to essentially fold non-meaningful variable values by merging adjoining *thin personas*, which are the personas that have the same locked variables and differing only by the *collapsible variables*. Each value for each variable within the set of *collapsible variables* is assigned a cost. Then, for each unique set of set locked variables, we seek to minimize the cost of the set of associated collapsible variables. We seek to optimize the cost function to the point where any further reduction in cost would cause a change in our set of locked variables. Our function returns the smallest value when the set of *collapsible variables* cannot be reduced further. Table 3 shows the *locked* and *collapsible variables*, with the values and cost of each of the *collapsible variables*.

To accomplish this, heuristically, we first define a set of variables that cannot be collapsed (*locked variables*). The remaining variables are, by definition, ones that can be collapsed (*collapsible variables*). We define all the behavioral variables and one demographic variable (Nationality) as *locked variables* to maintain personas rooted in behavioral variation, i.e., flight patterns in the data. We define two variables as *collapsible variables*: Age with 7 possible values, and Gender, with two possible values because demographic variation with these variables is considered less important than behavioral, as different demographic segments may be interested in the same flight destinations (apart from nationality, which is a strong predictor of flight behavior. Our goal is to collapse Age and Gender into the smallest set of values without causing a change in the locked values. Given that the Age variable has 7 possible values and the Gender variable has 2 possible

Table 5: Sample Demographic Groups with Corresponding Flight Pattern.

Nationality	*Age*	*Gender*	*Flight Pattern (set of 10 destinations)*
Germany	18_24	Male	1
Germany	25_34	Female	1
Germany	25_34	Male	1
Germany	35_44	Male	1

Table 6: Collapsed Persona Numbers with Occurrences and Percentage of Reduction.

Collapsible Personas	*Occurrence*	*% of Personas*	*Reduction by Group*
14	1	0.20%	93%
11	2	0.41%	71%
10	1	0.20%	64%
9	7	1.42%	57%
8	4	0.81%	50%
7	6	1.22%	43%
6	12	2.43%	36%
5	33	6.69%	29%
4	31	6.29%	21%
3	54	10.95%	14%
2	110	22.31%	7%
1	232	47.06%	0%
	493	100.00%	69%

values, there is a maximum of 14 *thin personas* that can be collapsed into each demographic-behavioral set.

5 RESULTS

We first applied our matrix factorization to approximately 26 million records using the flight booking destinations from customers over a two-year period. The result of this procedure is the identification of sets of ten flight destinations. We select the ten top most frequently occurring flight patterns, representing more than 90% of the original dataset. For space considerations and confidentiality concerns of the company, we refer to them as Flight Pattern 1, Flight Pattern 2, … Flight Pattern 10. These ten flight patterns are associated with one or more demographic groups, defined as (Age, Gender, Nationality). In total there are 1,598 *thin personas* composed of a set of demographics and one of the unique flight patterns. However, some of these *thin personas* are not practically distinct. For example, the four segments presented in Table 5. From just examining these *thin personas* in Table 5, we can see that there is no practical difference among these four personas, as a company could, and probably should, treat these personas as one for marketing, advertising, or website design purposes. Therefore, we can collapse these four personas into one persona, Germany_18-44_Male/Female, as the key components (Flight Pattern, other behavioral attributes we considered important, and Nationality are the same).

After applying the algorithm described above, we generated 493 *thin personas*, a nearly 69% reduction in the number of personas via a collapsing of Age and Gender variables. Examples of the collapsed *thin personas* are shown in Figure 1. From these, we can enrich to develop full persona profiles via adding actual photos and names. The effects of the collapse are shown in Table 6. We see that more than half (53%) of the *thin personas* are combined with at least one other *thin personas*. Approximately 69% of the *thin personas* were collapsed with one or two other *thin personas*. However, there was a long tail (31%) of *thin personas* that were collapsed with three or more *thin personas*. There was one set where all 14 segments were collapsed into one *thin personas*. In aggregate, 4.26% of the *thin personas* sets were reduced by more than half, 48.68% reduced by less than half, and 47.06% had no reduction, meaning that each *thin personas* in the demographic-behavioral set were somehow unique. Concerning what was most collapsible, we find that 46% of the collapsing was due to the Age variable, 23% due to the Gender variable. This is in line with our factor analysis, showing the Age was a poor predictor, while Gender was a weak predictor. 18% of the collapsing occurred with both the Age-Gender variables. We also see from Figure 2 that there is a rough correlation between the number of users in the demographic groups and the number of segments that can be collapsed. In other words, the larger the number of users, the more overlap among Age and Gender and the less important these variables are for persona generation.

6 DISCUSSION AND IMPLICATIONS

In this research, we show that one can generate a large number of personas for a diverse user population, and then reduce this large set of personas to a minimal set while still representing all user segments in the

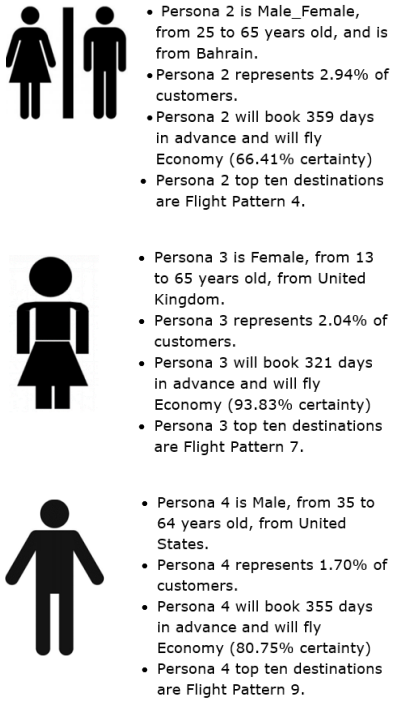

Figure 1: Four Exemplar Personas After Collapsing.

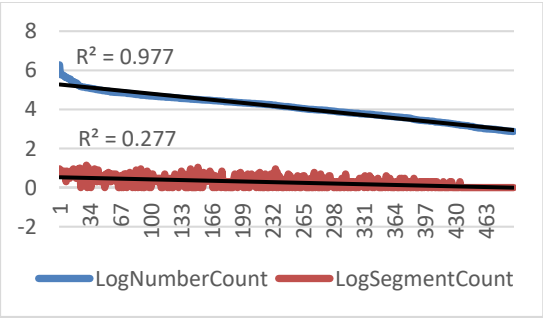

Figure 2: Log-log graph Illustrating the Correlation Between Number of Users and Collapsible Segments.

population. In fact, using behavioral online analytics data, persona creation can be automated [7]. This data-driven persona creation is a major shift from the way personas have traditionally been developed [6], and it offers benefits such as creating hybrid personas from both quantitative and qualitative data. This addresses a critical issue of combining big data and personas. The collapsed personas represent the user segments more effectively and efficiently, as there are no 'false' segments (i.e., segments are not behaviorally different from one or more other segments). Our approach ensures that this proper segmentation actually occurs without introducing a false distinction that does not actually impact the end system or product. The collapsing approach is also compatible with the best practice of persona creation according to behaviors that are more important for personas than the more superficial demographic traits. More empirical research is needed to evaluate how decision makers cope with a varying number of personas.

REFERENCES

[1] T. Adlin and J. Pruitt, *The Essential Persona Lifecycle: Your Guide to Building and Using Personas*: Morgan Kaufmann Publishers Inc., 2010.

[2] J. An, H. Kwak, S. Jung, J. Salminen, M. Admad, and B. Jansen, "Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data," *ACM Trans. Web*, vol. 12, pp. 1-26, 2018.

[3] M. F. Clarke, "The Work of Mad Men that Makes the Methods of Math Men Work: Practically Occasioned Segment Design," presented at the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea, 2015.

[4] S. Dolnicar, B. Grün, and F. Leisch, *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*, 1 ed. Singapore: Springer, 2018.

[5] S. Faily and I. Flechais, "Persona cases: a technique for grounding personas," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 2011.

[6] K. Goodwin and A. Cooper, *Designing for the Digital Age: How to Create Human-Centered Products and Services*. Indianapolis, IN: Wiley, 2009.

[7] S. Jung, J. An, H. Kwak, M. Ahmad, L. Nielsen, and B. J. Jansen, "Persona Generation from Aggregated Social Media Data," in *ACM Conference on Human Factors in Computing Systems 2017 (CHI2017)*, Denver, CO, 2017, pp. 1748-1755.

[8] L. Laporte, K. Slegers, and D. D. Grooff, "Using correspondence analysis to monitor the persona segmentation process," presented at the Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design, Copenhagen, Denmark, 2012.

[9] Z. Liu and Bernard J. Jansen, "Questioner or question: Predicting the response rate in social question and answering on Sina Weibo," *Information Processing & Managemen*, vol. 54, 2018.

[10] L. Nielsen, *Personas - User Focused Design*. London: Springer-Verlag, 2013.

[11] J. Pruitt and J. Grudin, "Personas: Practice and Theory," in *Proceedings of the 2003 Conference on Designing for User Experiences*, San Francisco, California, 2003, pp. 1-15.

[12] J. Salminen, S. Şengün, H. Kwak, B. J. Jansen, J. An, S. Jung, *et al.*, "From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas," *First Monday*, vol. 23, p. http://firstmonday.org/ojs/index.php/fm/article/view/8415, 2018.

[13] J. Salminen, S. Şengün, H. Kwak, B. J. Jansen, J. An, S. G. Jung, *et al.*, "Generating Cultural Personas From Social Data: A Perspective of Middle Eastern Users," in *The Fourth International Symposium on Social Networks Analysis, Management and Security (SNAMS-2017)*, Prague, Czech Republic, 2017, pp. 120-125.

[14] P. Sánchez and A. Bellogíns, "Building user profiles based on sequences for content and collaborative filtering," *Information Processing & Management*, vol. 56, pp. 192-211, 2019.

[15] P. Zerbino, D. Aloini, R. Dulmin, and V. Mininno, "Big Data-enabled Customer Relationship Management: A holistic approach," *Information Processing & Management*, vol. 54, pp. 818–846, 2018.