Persona Generation from Aggregated Social Media Data

Soon-Gyo Jung
Qatar Computing Research Institute, Hamad Bin Khalifa University
sjung@hbku.edu.qa

Moeed Ahmad
Strategy & Development Division
Al Jazeera Media Network
ahmadmo@aljazeera.net

Jisun An
Qatar Computing Research Institute, Hamad Bin Khalifa University
jan@hbku.edu.qa

Lene Nielsen
IT University of Copenhagen
lene@itu.dk

Haewoon Kwak
Qatar Computing Research Institute, Hamad Bin Khalifa University
hkwak@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute, Hamad Bin Khalifa University
jjansen@acm.org

CHI'17 Extended Abstracts, May 06-11, 2017, Denver, CO, USA
ACM 978-1-4503-4656-6/17/05.
http://dx.doi.org/10.1145/3027063.3053120

Abstract
We develop a methodology for persona generation using real time social media data for the distribution of products via online platforms. From a large social media account containing more than 30 million interactions from users from 181 countries engaging with more than 4,200 digital products produced by a global media corporation, we demonstrate that our methodology can first identify both distinct and impactful user segments and then create persona descriptions by automatically adding pertinent features, such as names, photos, and personal attributes. We validate our approach by implementing the methodology into an actual working system that leverages large scale online user data for generation of persona descriptions. We present the overall methodological approach, data analysis process, and system development. Findings show this method can develop believable personas representing real groups of people using real-time online user data. Results have implications for those distributing products via online platforms.

Author Keywords
Persona; User Experience Research; User Analytics

ACM Classification Keywords
H.5.2. [Information Interfaces and Presentation] User Interfaces - Theory and methods, User-centered design

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. Can be based on theories from sociology.
Social Media Platform Technologies that provide for the creating and sharing of information and other forms of expression via virtual communities and networks.
Social Media Analytics Practice of gathering data from social media platforms and analyzing that data to make informed decisions.

Table 1. Key Constructs and Definitions.

Introduction

Personas are representations of segments of actual users, presented as an imaginary person. It is used in IT-development, system design, and marketing. The final artifact is typically a persona description embodying attributes of the user segment that the fictionalized person represents. Personas are a continuation of efforts from a variety of domains for identifying, assessing, and constructing groups of people (i.e., users, customers, audience, or market segments) in order to optimize some performance metric (i.e. speed of task or ease of use). Personas are reportedly well integrated into current design processes [1-5] for both long and short term projects [4].

Although personas have claimed benefits beyond what data analytics by itself can provide for identifying user segments [1-4, 6-12], there are still questions concerning the value of personas [12-16], most notably the challenge to develop [12, 14, 17]. Creating personas is not a cheap, easy, or quick process, as the construction has historically involved ethnography studies or focus groups. As these are one time data collection events, the personas created can also become quickly outdated. Without updated data, designers have no confirmation whether the personas are representative of their current target users. These limitations are especially acute for those that distribute products via major online platforms. With a potential audience in the millions or billions, traditional ethnography methods do not scale and are cost prohibitive. These limitations are the motivations for our research, in which we propose, develop, and implement an approach for leveraging privacy-preserving aggregated data of user interactions with products from online platforms and then enriching the analysis results with descriptive attributes to generate persona descriptions. Our approach can be a stand along method or complementary with qualitative means of persona development.

Review of Literature

Although the assumption is that personas are developed from data representing real users [18], this has not always been the case, as the data is sometimes time consuming and expensive to gather [5]. When real data is used, personas are typically developed from fieldwork, such as user interviews, direct observations, etc. [8, 19]. The analysis methods employed depend in large part on the user data availability, and a critique is that many personas are not based on sizeable quantities of first-hand user data, or there is not enough data for quantitative methods [20]. Therefore, the creation of personas from a quantitative, data-driven approach based on actual first hand user data of sizeable quantities has remained an open research question [21], with few efforts reported [20].

Research Objective

Our objective is to develop a methodology for mining aggregated large scale user data from major social media platforms (e.g., Facebook, Twitter, YouTube, etc.) in order to identify user segments, both distinct and impactful, and then automatically generate personas with realistic persona descriptions that represent these key user segments.

This research is novel in several respects. First, it is one of, if not the, first systems to use real user data to automate the persona generation process. Second, the data from such platforms is already aggregated, unlike prior work in identifying audience segments [22, 23],

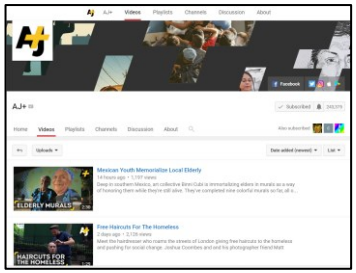

Figure 1: AJ+ YouTube Channel with hundreds of thousands of followers for their thousands of videos.

Figure 2: Example AJ+ content, which receives thousands to millions of views.

Figure 3: AJ+ YouTube content generates not only views but lively conversations in the comments.

meaning that we must introduce techniques to de-aggregate it. Third, the increase in the number of needed personas. Typically, persona generation has focused on a small number of personas, three to six. While appropriate for a traditional system environment, it is not appropriate for products that may be viewed by millions to billions of users. Beyond our work [24-26], we could locate no prior research concerning generating personas for those who distribute their products via major platforms. This research builds on our prior work [24-26] via a more robust personas development.

Data Collection

To demonstrate the feasibility of our approach, we leverage user data from AJ+, an online news channel from Al Jazeera Media Network. A common goal for many organizations is to increase digital content consumption and enhance the platform's facilitation of digital content interaction. Therefore, proper understanding of audiences is critically important, which online user data can provide [27, 28].

In this research, we collaborate with AJ+ for both data collection (see Table 2) and analysis. We focus on the AJ+ YouTube channel as the source of the aggregated audience statistics, which we use as a proof-of-concept for our persona generation methodology (see Figures 1 through 3).

The YouTube analytics API provides, for each AJ+ video, various user profile data, (e.g., gender, age, country location, and which site the user comes from), although at an aggregate level. Via the API, we collect the detailed breakdown of views by country, gender, and age group. We focus here on view counts and viewer percentage due to their high volumes. We also note that this detailed breakdown data is accessible only to an owner of a YouTube channel (i.e., AJ+). In summary, we collect data from 4,320 videos uploaded from June 13, 2014 to July 27, 2016. These videos had more than 30 million views via users from 181 countries at the time of the study (see Figure 4).

Methodology for Automatic Persona Generation from Social Media Data

Approach Formulation

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

a) identifying distinct user interaction patterns from the data,
b) linking these distinct user interaction patterns to user demographic groups,
c) identifying impactful user demographic groups from the data,
d) creating skeletal personas via demographic attributes, and
e) enriching these skeletal personas to create rich personas description.

We first develop a matrix representing users' interaction with the online products. We denote by V the g×c matrix of g user groups (G_1, G_2, ..., G_g) and c contents (C_1, C_2,...,C_c). The element of the matrix V, V_{ij}, is any statistic that represents the interaction of user group G_i for content C_j. With this matrix as the basis, we can identify first distinct user behavior patterns (which can be patterns of any set of user touch points) and then the impactful user segments from this set of distinct user patterns.

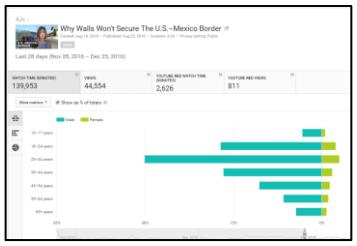

Figure 4: Snapshot of the AJ+ YouTube Analytics interface, from which we access the data via the YouTube API.

Platform: YouTube
Channel: AJ+
Subscribers: ~180,000 at time of study
Period: 13 June through 27 July 2014
Products: 4,320 videos at time of study
Interactions: ~30 million at time of study
Localities: users from 181 countries at time of study

Table 2. Key Statistics of AJ+ YouTube Data Set. A sizeable data collection effort with thousands of content products, tens of millions of interactions, and a heterogeneous user based.

Identification of Distinct User Segments

Once we have the matrix V, discovering the underlying latent factors, the user interaction patterns, that will become the basis of the personas is the next step. There are several ways to decompose a given matrix; for this proof-of-concept, we used non-negative matrix factorization (NMF) [29] (see Figure 5). Compared to a simple clustering of user groups, NMF has advantages in that it can find multiple behavioral patterns even from a single group.

The matrix decomposition works as following:

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

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.

Figure 5: Brief overview of NMF, which identifies distinct interaction patterns and then impactful demographic segments, which form the basis for the resulting personas.

Matrix H shows a set of common content consumption patterns signified by a linear combination of content, representing the set of distinct user behavior patterns (see Figure 5).

Identification of Impactful User Segments

We next focus on W, shown in Figure 5. A row in W represents how each user group can be characterized by different consumption patterns. A column in 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 the most impactful user group for that corresponding pattern.

Incorporating Attributes into a Persona Description

The result of NMF is a set of skeleton personas, which we turn into rich personas by adding personality attributes, as outlined in the following.

Persona Demographics

We determine the demographics of the representative user groups. Depending on how the user groups are defined in V, the most efficient way is to use the data broken-out by demographics when building V. If we build V where a row maps to a group defined as <age group, gender, country>, then it is trivial to find representative demographics of a persona. Social media analytic tools often provide user statistics in a format that we can leverage for a persona profile descriptive snippet (see Figure 6).

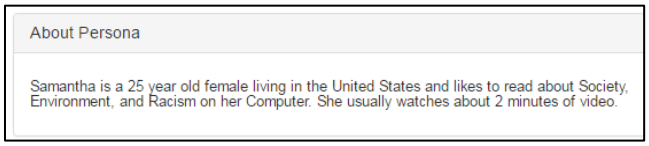

Figure 6: Resulting persona snippet for the personas description, with the demographic information from the NMF, augmented with a gender and temporally appropriate name, topical interests, and key user behavior.

In our data collection site, for example, YouTube provides demographic classification of 2 genders x 7 age groupings x 198 countries (2,772 possible demographic grouping) per video, which is the floor of

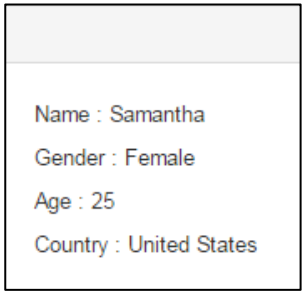

Figure 7: Augmented demographic attributes with a gender and temporally appropriate name automatically assigned for the persona description heading.

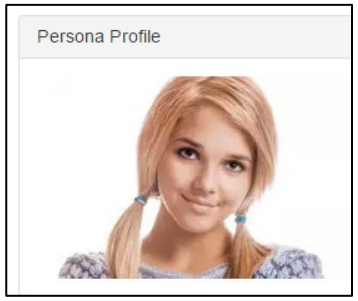

Figure 8: Gender and temporally appropriate photo automatically assigned. For system development, we purchased the copyrights to nearly 4,000 photos in order to ensure all demographics sets are covered and each persona has a unique photo.

possible demographic personas that that can be addressed.

Persona Name

To generate a name for a persona, we build a dictionary of names by collecting popular names by gender and year from the 181 countries. For example, there is information on the US and many other countries concerning the top 1,000 popular baby names for any year since 1879. Then, through <age group, gender, and country> of a representative group, we can automatically assign an age, gender, and ethnically appropriate name to a persona (see Figure 7).

Persona Photo

To assign a photo to a persona, we purchase copyrights to nearly 4,000 commercial stock 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, country, etc.> of a representative user segment, we can assign an appropriate photo to a persona (see Figure 8).

Other Personas Details

We can filter users based on topical interests (see Figure 9) by leveraging the collection of content viewed by that persona (see Figure 10). We are experimenting with various topically classifiers at the moment. Once identified, we can use this to link to other user online behaviors. For example, if a persona watches many videos about soccer, it is a reasonable assumption that users whose tweets are mainly about soccer in Twitter potentially have a similar content consumption pattern. We can leverage like demographics to get social media comments about the online products, which we can incorporate into the persona description (see Figure 11).

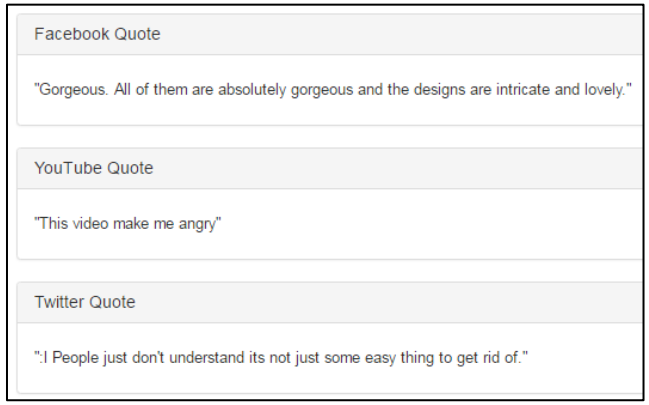

Figure 11: Once key demographic attributes and user behavior interaction patterns are identified, we pull comments representing those made by this 'persona'. These comments are pulled every time the persona description is displayed, so new interactions are shown by the system.

The result of this attribute incorporation are rich insightful persona descriptions, generated automatically, all from aggregated, privacy preserving social media data, as shown in Figure 12.

Discussion and Implications

In this research, we demonstrate several important results and implications concerning personas. Notably, the approach shows that one can use real time, aggregated online user data at scale in order to identify meaningful user segments and then automatically generate personas. As such, this research addresses a

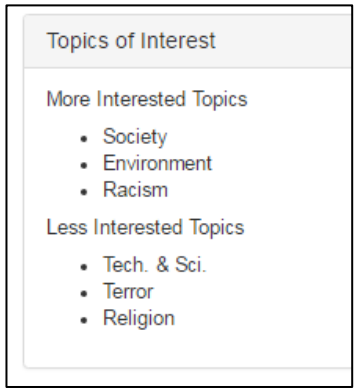

Figure 9: Example of topical interest classification for each persona. Topical interests are ranked from most to least interested.

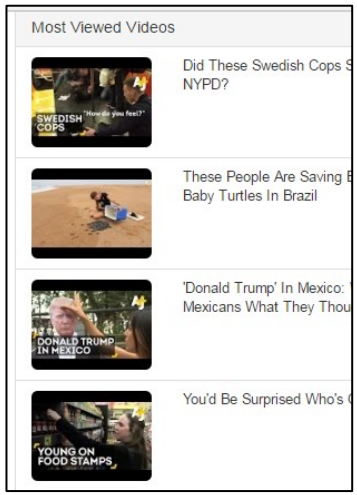

Figure 10: Example listing of top watched videos.

previously open question of investigation and advances persona research by presenting an approach to use data at scale and leverage this data continually to keep personas updated.

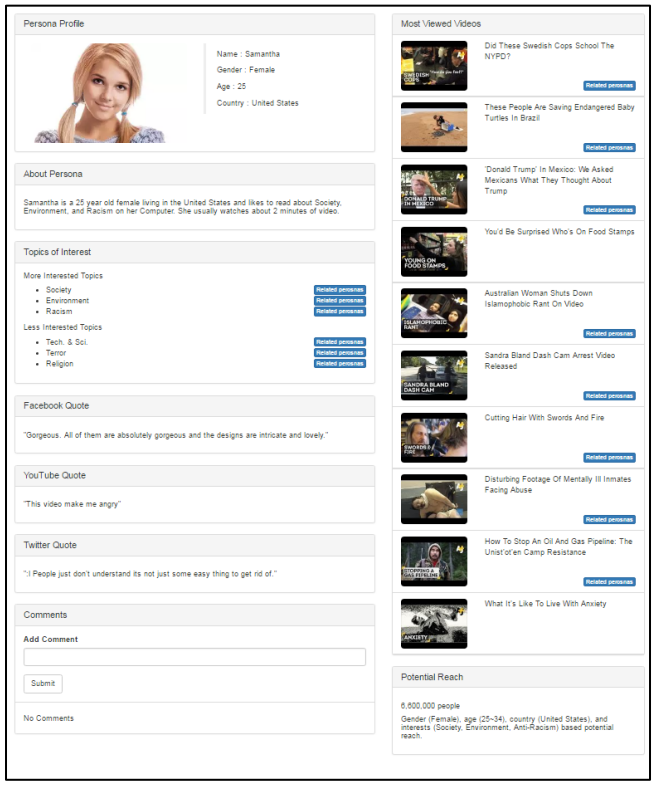

Figure 12: Rich persona description, with traditional components and persona attributes, all generated automatically from online social media data. Additionally, as the system is online, one can directly access the underlying data providing credence to the personas [30].

Also, our research focuses on the use of data from major online platforms that are, in most cases, aggregated. Prior work used individual user data, to which many content creators do not have access. Our research using NMF demonstrates that one can use this aggregate data to both identify distinct and impactful user segments and, then automatically generate rich personas. This research also focuses on the increasingly common situation of digital content creators that are distributing their content to an extremely large, heterogeneous user base via major online platforms, which is, or is becoming, the de facto technology of distribution.

There are several areas for future research. Given the reliance on streams of social media data, we could certainly implement direct access to the foundational user data for the content creators and also persona campaigns, as suggested by [30], where updates concerning the personas are continually sent to the product developers. This feature may be important, as it appears that designers like continued access to the actual user data, aside from the persona description itself, to aid them in their design decisions [30]. We are also investigating other approaches besides NMF to identify user segment.

Conclusion

In this research, we show 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 reflecting real people. Although focusing on digital content creators, our approach is flexible and resilient for application in a wide range of contexts, which we are exploring.

References

1. Pallavi Dharwada, Joel S. Greenstein, Anand K. Gramopadhye, and Steve J. Davis, "A Case Study on Use of Personas in Design and Development of an Audit Management System," in *Human Factors and Ergonomics Society Annual Meeting Proceedings*, Baltimore, Maryland, 2007, pp. 469-473.
2. Elina Eriksson, Henrik Artman, and Anna Swartling, "The Secret Life of a Persona: When the Personal Becomes Private," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 2013.
3. Erin Friess, "Personas and Decision Making in the Design Process: An Ethnographic Case Study," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA, 2012.
4. Tejinder Judge, Tara Matthews, and Steve Whittaker, "Comparing Collaboration and Individual Personas for the Design and Evaluation of Collaboration Software," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA, 2012.
5. Lene Nielsen and Kira Storgaard Hansen, "Personas is Applicable: A Study on the Use of Personas in Denmark," presented at the Proceedings of the 32nd annual ACM conference on Human factors in computing systems, Toronto, Ontario, Canada, 2014.
6. Tamara Adlin and John Pruitt, *The Essential Persona Lifecycle: Your Guide to Building and Using Personas*: Morgan Kaufmann Publishers Inc., 2010.
7. Hugh Beyer and Karen Holtzblatt, *Contextual Design: Defining Customer-centered Systems*: Morgan Kaufmann Publishers Inc., 1998.
8. Kim Goodwin and Alan Cooper, *Designing for the Digital Age: How to Create Human-Centered Products and Services*. Indianapolis, IN: Wiley, 2009.
9. Adrienne L. Massanari, "Designing for Imaginary Friends: Information Architecture, Personas, and the Politics of User-Centered Design," *New Media & Society,* vol. 12, pp. 401-416, 2010.
10. Tomasz Miaskiewicz, Susan Jung Grant, and Kenneth A Kozar, "A Preliminary Examination of Using Personas to Enhance User-Centered Design," in *AMCIS 2009 Proceedings*, 2009, p. Article 697.
11. John Pruitt and Jonathan Grudin, "Personas: Practice and Theory," presented at the Proceedings of the 2003 conference on Designing for user experiences, San Francisco, California, 2003.
12. K. Ronkko, "An Empirical Study Demonstrating How Different Design Constraints, Project Organization and Contexts Limited the Utility of Personas," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 2005, pp. 220a-220a.
13. Åsa Blomquist and Mattias Arvola, "Personas in Action: Ethnography in an Interaction Design Team," presented at the Proceedings of the second Nordic conference on Human-computer interaction, Aarhus, Denmark, 2002.
14. Christopher N. Chapman and Russell P. Milham, "The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method," in *Human Factors and Ergonomics Society Annual Meeting*, San Francisco, CA, 2006, pp. 634-636.
15. Steve Portigal. (2008, 29 December). *Persona Non Grata*. Available: http://www.portigal.com/wp-content/uploads/2008/01/Portigal-Consulting-White-Paper-Persona-Non-Grata.pdf}
16. Kari Rönkkö, Mats Hellman, Britta Kilander, and Yvonne Dittrich, "Personas is not Applicable: Local

Remedies Interpreted in a Wider Context,," presented at the Proceedings of the eighth conference on Participatory design: Artful integration: interweaving media, materials and practices - Volume 1, Toronto, Ontario, Canada, 2004.
17. Rósa Gu!jónsdóttir and Sinna Lindquist, "Personas and Scenarios: Design Tool or a Communication Device," in *8th International Conference on Cooperative Systems (COOP'08)*, Carry-le-Rouet, France, 2008, pp. 165-176.
18. John Pruitt and Tamara Adlin, *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*: Morgan Kaufmann Publishers Inc., 2005.
19. Alan Cooper, *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity (2nd Edition)*: Pearson Higher Education, 2004.
20. Jennifer McGinn and Nalini Kotamraju, "Data-driven Persona Development," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, 2008.
21. Kerry Rodden, Hilary Hutchinson, and Xin Fu, "Measuring the User Experience on a Large Scale: User-centered Metrics for Web Applications," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA, 2010.
22. Bernard J. Jansen, Kate Sobel, and Geoff Cook, "Classifying Ecommerce Information Sharing Behaviour by Youths on Social Networking Sites," *Journal of Information Science,* vol. 37, pp. 120-136, 2011.
23. Xiang Zhang, Hans-Frederick Brown, and Anil Shankar, "Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry," presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, Santa Clara, California, USA, 2016.
24. J. An, H. Cho, H. Kwak, M. Z. Hassen, and B. J. Jansen, "Towards Automatic Persona Generation Using Social Media," in *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, 2016, pp. 206-211.
25. 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 (AICCSA2016)*, Agidar, Morocco, 2016.
26. H. Kwak, J. An, and B. J. Jansen, "Automatic Generation of Personas Using YouTube Social Media Data," in *Hawaii International Conference on System Sciences (HICSS-50)*, Waikoloa, Hawaii, 2017, pp. 833-842.
27. E. Mao and J. Zhang, "What Drives Consumers to Click on Social Media Ads? The Roles of Content, Media, and Individual Factors," in *2015 48th Hawaii International Conference on System Sciences*, 2015, pp. 3405-3413.
28. G. Shuradze and H. T. Wagner, "Towards a Conceptualization of Data Analytics Capabilities," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 2016, pp. 5052-5064.
29. Daniel D. Lee and Sebastian H. Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature,* vol. 401, pp. 788-791, 1999.
30. Tara Matthews, Tejinder Judge, and Steve Whittaker, "How Do Designers and User Experience Professionals Actually Perceive and Use Personas?," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA, 2012.