Giving Faces to Data: Creating Data-Driven Personas from Personified Big Data

Soon-Gyo Jung, Joni Salminen, Bernard J. Jansen
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
{sjung, jsalminen, bjansen}@hbku.edu.qa

ABSTRACT

Creating personas from large amounts of online data is useful but difficult with manual methods. To address this difficulty, we present Automatic Persona Generation (APG), which is an implementation of a methodology for quantitatively generating data-driven personas from online social media data. APG is functional, and it is deployed with several organizations in multiple industry verticals. APG employs a scalable web front-end user interface and robust back-end database framework processing tens of millions of user interactions with tens of thousands of online digital products across multiple online platforms, including Facebook, Google Analytics, and YouTube. APG identifies audience segments that are both distinct and impactful for an organization to create persona profiles. APG enhances numerical social media data with relevant human attributes, such as names, photos, topics, etc. Here, we discuss the architecture development and central system features. Overall, APG can benefit organizations distributing content via online platforms or with online content that relates to commercial products. APG is unique in its algorithmic approach to processing social media data for customer insights. APG can be found online at https://persona.qcri.org.

CCS CONCEPTS

• H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces; K.4.m. Computers and society: Miscellaneous.

KEYWORDS

Personas, data-driven personas, persona development

ACM Reference format:

Jung, S. G., Salminen, J., and Jansen, B. J. (2020) Giving Faces to Data: Creating Data-Driven Personas from Personified Big Data. *In Proceedings of Intelligent User Interfaces (ACM IUI2020).* 17-20 March. Cagliari, Italy.

1 Introduction

With massive potential reach, data from social media platforms can provide insights for user analysis. Working with this data, however, has certain disadvantages, including its high volume. Personas are one approach to simplifying this user data and give a human face to these sterile numbers. The concept of personas is employed in various domains, e.g., system development, software design, advertising, and marketing for describing and communicating about core users, customers, or audiences. One can view of personas as conceptual shortcuts for presenting the set of people that are (or can be) users, audience members, or customers. When employed, personas can aid organizations to define strategic goals, develop targeted products, and improve customer-centered operations. Many designs, product, marketing, content, and advertising development processes can be enhanced by integrating personas.

Automatic Persona Generation (APG) system uses aggregated, privacy-preserving data of user interactions with digital content that is posted on online social media and website platforms. The APG system collects, processes, and decomposes this authentic user data, and then enriches the initial results to include descriptive attributes to produce truly data-driven persona profiles. APG can generate personas from millions of user interactions within a matter of hours, far quicker than traditional persona creation techniques that rely on manual data collection and provide a human face for understanding users that analytics alone cannot.

We present and discuss an overview of APG, focusing on the architecture and development approach. Then, we highlight the central APG features, with other non-central features presented in the actual demonstration. Also, we present some potential use cases, system uniqueness, and on-going commercial efforts.

2 System Overview

The APG system leverages an established and scalable architectural structure that is also robust in terms of data expansion by employing: (a) a web framework, Flask, to support front-end application development, creation of services, and for application programming interfaces (APIs), (b) an open-source database, PostgreSQL, for the back-end data storage and data processing, and (c) Python libraries such as Pandas and scikit-learn for data processing and analysis [1].

APG links to and accesses the specific online social media platforms, (e.g., Facebook, Google Analytics, YouTube) via the API of each analytics platform, given the account holder's permission. The typical user data provided by these platforms include the demographic variables of gender, age, and country of the person, provided at an aggregated group level. Via the APIs, APG collects the detailed interactions of users with each of the online content pieces on the corresponding platform. This data is available to

IUI '20 Companion, March 17–20, 2020, Cagliari, Italy

ACM ISBN 978-1-4503-7513-9/20/03.
https://doi.org/10.1145/3379336.3381465

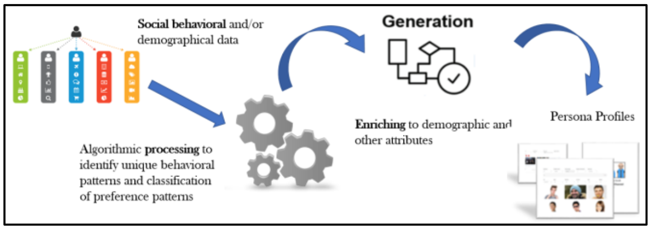

Social behavioral and/or demographical data
Generation
Algorithmic processing to identify unique behavioral patterns and classification of preference patterns
Enriching to demographic and other attributes
Persona Profiles

Figure 1: APG data and processing flowchart from server configuration to data collection and persona generation.

only owners of a particular social media channel (e.g., YouTube channel) and not available to the general public.

APG applies a sequential methodology [1, 3]approach to automatically generate personas from this social media data, consisting of:

- calculating users' distinct interaction patterns with content,
- linking distinct patterns to specific demographic groups,
- identifying an impactful demographic group for each of the distinct patterns
- creating thin personas via the combination of behavioral and demographic attributes,
- enriching these thin personas to generate complete persona profiles.

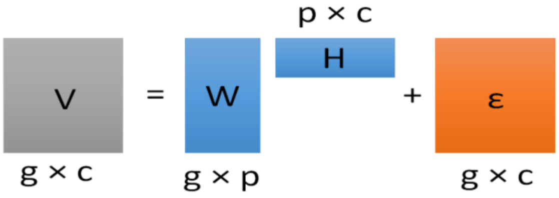


Figure 2: Matrix decomposition using NMF. Matrix V is decomposed into W and H. g denotes demographic groups in the dataset, c denotes product units, p is the number of latent behaviors of demographic groups over product units, and ε is the error term.

For the algorithmic implementation of this process, we apply Non-negative Matric Factorization (NMF) for identifying latent customer interaction (see Figure 2). NMF is particularly intended for reducing the dimensionality of large datasets by discerning latent factors [5]. From this foundational 'persona', the system implements a process of enrichment by adding an appropriate name, picture, social media quotes, and related demographic attributes (e.g., marital status, educational level, occupation, etc.) via querying the Facebook API. The result is a set of organizational persona profiles representing the customer population segments. The personas are shown to end users (i.e., the people from the organization whose data is used for the persona generation) via the online system running on Flask[1], an open-source Python web framework (see Figure 3). APG is a real, fully functional system deployed with real client organizations, and a demo available online[2].

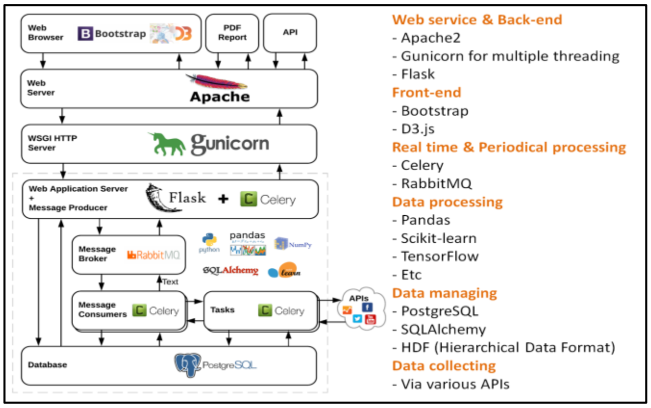
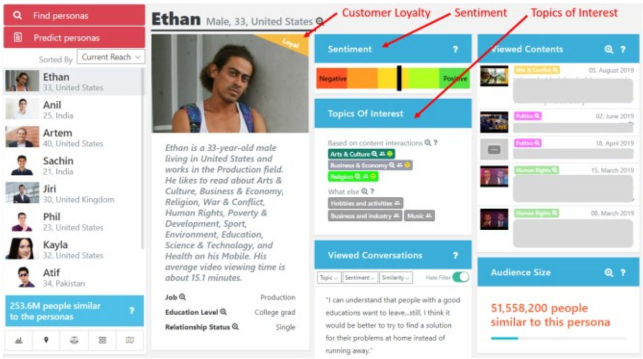


Figure 3: APG Infrastructure.

A detailed technical explanation of APG's system infrastructure is found in Jung et al. [2]. The persona profile (see Figure 4) has nearly all of the typical persona attributes. Additionally, data-driven personas, by relying on regular data collection intervals, can enrich the traditional persona profile with additional elements such as (a) customer loyalty, (b) sentiment analysis [4], and (c) topics of interest (see Figure 4). Also, the persona profile is interactive and functions as the interface to the additional level of analytics and user data.



Ethan Male, 33, United States

Ethan	33, United States
Anil	25, India
Artem	40, United States
Sachin	21, India
Jiri	30, United Kingdom
Phil	23, United States
Kayla	32, United States
Atif	34, Pakistan

253.6M people similar to the personas

Ethan is a 33-year-old male living in United States and works in the Production field. He likes to read about Arts & Culture, Business & Economy, Religion, War & Conflict, Human Rights, Poverty & Development, Sport, Environment, Education, Science & Technology, and Health on his Mobile. His average video viewing time is about 15.1 minutes.

Job	Production
Education Level	College grad
Relationship Status	Single

"I can understand that people with a good educations want to leave...still, I think it would be better to try to find a solution for their problems at home instead of running away."

05. August 2019
02. June 2019
18. April 2019
15. March 2019
08. March 2019

51,558,200 people similar to this persona

Figure 4: The APG personas listing (left of screen) and a personas profile displayed. An APG Persona Profile contains the standard persona profile attributes, plus with direct access to the underlying data used for creation.

REFERENCES

[1] An, J. et al. 2018. Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Transactions on the Web (TWEB).* 12, 4 (2018), Article No. 27. DOI:https://doi.org/10.1145/3265986.

[2] Jung, S. et al. 2018. Automatically Conceptualizing Social Media Analytics Data via Personas. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2018)* (San Francisco, California, USA, Jun. 2018).

[3] Jung, S. et al. 2017. Persona Generation from Aggregated Social Media Data. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA, 2017), 1748–1755.

[4] Kumar, A. et al. 2020. Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management.* 57, 1 (2020), 102141. DOI:https://doi.org/10.1016/j.ipm.2019.102141.

[5] Lee, D.D. and Seung, H.S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature.* 401, 6755 (Oct. 1999), 788–791. DOI:https://doi.org/10.1038/44565.

[1] http://flask.pocoo.org

[2] https://persona.qcri.org