Chapter 8
Automatic Persona Generation for Online Content Creators: Conceptual Rationale and a Research Agenda

Joni Salminen, Bernard J. Jansen, Jisun An, Haewoon Kwak and Soon-Gyo Jung

8.1 Introduction

As the quantity of social and online analytics data has drastically increased, a wide variety of methods are deployed to make sense of this data, typically via computational and algorithmic approaches. However, in many cases, these approaches trade one form of complexity for another by ignoring the principles of human cognitive processing. In this perspective manuscript, we propose an approach of employing Personas as an alternative form of making large volumes of online user analytics information useful to end users of the user and customer analytics, with results applicable in software development, business sectors, communication industry, and other domains where understanding online user behavior is deemed important. Toward this end, we have developed a system that automatically generates data-driven Personas from social media and online analytics data, capable of handling hundreds of millions of user interactions from tens of thousands of pieces of content on YouTube, Facebook and Google Analytics, while retaining the privacy of individual users of those channels. Our approach (1) identifies and prioritizes user segments by their online behavior, (2) associates the segments with demographic data, and (3) creates rich Persona profiles by dynamically adding characteristics, such as names, photos, and descriptive quotes. This chapter characterizes the currently open research problems in automatic Persona generation, such as de-aggregation of data, cross-platform data mapping, filtering of toxic comments, and choosing the right information content according to end-user needs. Addressing these problems requires the use of state-of-the-art techniques of computer and information science within one system and benefits greatly from inter-disciplinary collaboration. Overall, the research agenda set in this work aims at achieving the vision for automatic user profiling using diverse

J. Salminen (✉) · B. J. Jansen · J. An · H. Kwak · S.-G. Jung
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

J. Salminen
Turku School of Economics, Turku, Finland

L. Nielsen, *Personas - User Focused Design*, Human–Computer Interaction Series,
https://doi.org/10.1007/978-1-4471-7427-1_8

online and social media platforms and advanced data processing methods for the end goal of making complex analytics data more useful for human decision makers, especially those working with online content.

8.1.1 State of the Art in Data-Driven Personas

Overall, Personas are classified into three categories based on their usage of data: (a) Personas based solely on data, (b) Personas based on data but with considerable fictive elements, and (c) entirely fictive Personas created without data (Matthews et al. 2012). When real data is used, Personas are typically developed using ethnographic fieldwork and/or user interviews (Cooper 2004; Goodwin 2011; Pruitt and Grudin 2003). A major critique of Personas created manually is that they are often based on a small volume of user data, not enough to apply quantitative methods (Chapman and Milham 2006). Therefore, creating data-driven Personas based on behavioral data in large quantities has remained an open research question, with a limited number of efforts reported in the literature (McGinn and Kotamraju 2008).

While prior work on data-driven Personas remains scarce, there is more research on automated user profiling. For example, Guo et al. (2016) utilized social media data to develop credit risk profiles. Jansen et al. (2011) used the data from 35,000 social media users to cluster them based on content sharing patterns. However, the researchers did not generate Personas, but only assigned descriptive names to each cluster. In another work, Zhang et al. (2016) analyzed clickstream data and used hierarchical clustering to identify 10 common click flows. They generated five user Personas, giving them descriptive names. However, the information displayed in this study was limited in scope.

To generate more complete Personas, there is a new line of works attempting to fully automate Persona generation (An et al. 2016b, 2017; Jung et al. 2017; Kwak et al. 2017). An et al. (2016a) experimented with k-means clustering to generate Personas from social media data. However, that approach used individual-level data that is expensive or difficult to collect and has privacy concerns. In addition, Kwak et al. (2017) found that a single demographic group must fall into one Persona, while, in reality, several behavioral patterns can be found within one demographic group, as people in the same demographic group often behave in different ways. More precisely, the topical interests of the audience vary (An et al. 2017). Matrix factorization, briefly explained in this work, represents the best known solution to automatic Persona generation the authors are aware of (Kwak et al. 2017). However, many other sub-problems for creating credible, consistent and complete Personas remain to be solved (see Table 8.3).

Despite the progress made so far in automatic Persona generation, much remains to be studied and resolved. Automatic Persona generation is a particularly complex field due to the multitude of required computational approaches in solving Persona-generation issues, and due to the interdisciplinary nature of information selection, validation, and end-user needs. We expect the research agenda laid out here to carry

Table 8.1 Persona evaluation techniques

Technique	Description	Examples
Case studies	Conducting case studies within organizations to record the experiences of use, usefulness and impact of Personas for end users and decision makers	Rönkkö et al. (2004), Rönkkö (2005), Friess (2012), Jansen et al. (2017)
Stability analysis	Analyzing how stable data-driven Personas remain over time. Rapidly changing Personas would potentially indicate methodological problems	An et al. (2017)
Prediction	Analyzing how well different content is correlating with different Personas (when content interaction patterns are discriminators for Personas)	An et al. (2017)
Survey	Measuring end users' perception of Personas: are the created Personas perceived as credible, complete and consistent	Salminen et al. (2018a)

over to several years of active research, with the potential to open several new avenues of inquiry in multiple sub-fields.

8.1.2 Evaluation of Personas

Evaluation is a widely reported challenge in Persona research (Chapman and Milham 2006; Matthews et al. 2012; Miaskiewicz et al. 2008). Table 8.1 presents some existing evaluation techniques. Most commonly, Personas are evaluated in case studies. Their quantitative evaluation is rare, although there are some exceptions (An et al. 2016b; Chapman et al. 2008). Recently, survey-based validation has been proposed (Salminen et al. 2018).

8.2 Automatic Generation of Personas

This chapter proposes a research roadmap for the automatic generation of Personas from the large volumes of user analytics data available on the major social media and online analytics platforms, such as YouTube Analytics, Facebook Insights, and Google Analytics. The automatic generation of Personas addresses four major challenges of conventional Persona generation: (1) the slow and expensive process of manual Persona creation, (2) the difficulty of employing complex user analytics data for practical decision-making due to innate limitations of people to deal with numbers (Tversky and Kahneman 1974), (3) the "staling Persona" problem, meaning that manually created Personas do not automatically update but require repeating the data collection and analysis process, whereas automatic Persona generation does not

suffer from the same limitations, and, finally, the (4) concerns of representativeness of qualitative data collection (Chapman and Milham 2006). Overall, the development of a system and methodology for automatic Persona generation from online analytics data has special value for professionals working with online content. These include, for example, social media and community managers, news producers, content editors, bloggers and vloggers, and inbound marketers. Despite the enhanced availability of social media and online analytics data (or perhaps precisely because of it), these decision makers working with online content are struggling to turn data into insights (Salminen et al. 2017; Agarwal and Dhar 2014; Gandomi and Haider 2015; Jansen et al. 2017).

The automatic generation of Personas addresses the aforementioned issues via development of data-driven Persona profiles that present user information in a digestible and usable manner to which most humans can intuitively relate (Nielsen and Storgaard Hansen 2014), and that are based on millions of behavioral content interactions on the Web (Kwak et al. 2017). Additionally, these Personas can be created rapidly and updated easily, unlike manually created Personas that are cumbersome to develop, and they preserve the privacy of individual users due to the fact that the method leverages aggregated instead of individual-level data (An et al. 2016a, b; An et al. 2017). However, to accomplish the objective of fully automated Persona profiles, there remain several open research questions. The question relates to the broader field of computer science, and more specifically to user segmentation and profiling, natural language processing, topic modeling, information science, human-computer interaction (HCI), and Web analytics, among other areas. In addition, application of Personas in real usage scenarios and industry settings requires close collaboration with researchers and practitioners with topical domain expertise in those industries. It is our firm belief that the open research issues need to be addressed before automatically generated Personas can realize their full value and potential in use.

With this chapter, we aim to achieve the following objectives: (1) to clarify the ultimate goal and vision of automatic Persona generation, which is complex and multidimensional, (2) to provide a thorough research agenda for data-driven Personas in the age of online analytics, and (3) to outline how addressing these research problems opens new investigative avenues for information science, HCI, and other disciplines under computer science and related fields. By research agenda, we refer to a collection of research problems that are (a) logically associated with a greater mission or objective (here the achievement of fully automatic Persona profiles), (b) unsolved generally or in the particular context (here in the context of automatic Persona generation), (c) intellectually interesting and challenging (which is the measure for subjective judgment), and (d) have a plausible possibility of resulting in useful outcomes, i.e. practical impact, alongside theoretical implications. It is our firm belief that the research questions and the overall agenda laid out here fulfills these requirements. Finally, we hope to inspire research along multiple fronts within related and applicable domains: Personas have been applied across a variety of domains, including but not limited to commerce and marketing, e-health and design. The manuscript represents a call for action to researchers who are working in these fields and are

interested in automated customer/user insights, encouraging them to contribute in the methodological and practical development of data-driven Personas using their own contexts to pursue this agenda.

While it is commonplace for academic research to provide justifications relative to individual research articles, the long-term agenda of research projects is rarely clarified and communicated beyond funding application. Illuminating these aspects and rationales behind the research in the form of a reflective meta-discussion, we argue, can be beneficial for both the individuals engaged in research teams and the academic community as a whole. In the following sections, we show the need for automatic generation of Personas, discuss how automatically generated Personas leverage existing strengths of current approaches, outline the state-of-the-art research in this area, and then proceed to presenting outstanding research issues. As the objective of this manuscript relates to laying out a research agenda, we restrain from detailed technical descriptions—the details for automatic Persona generation are reported in other works (An et al. 2016b, 2017; Kwak et al. 2017). Instead, we give a general overview of the process of automatic Persona generation, briefly describing the steps of data collection, data processing, Persona generation, and interaction of the system with the end users. We then focus on laying out the agenda for on-going and future research.

8.2.1 Need for Automatic Persona Generation

Current Drawbacks with Personas

Although Personas have been shown to be useful in work environments where customer insights are a necessity (Blomquist and Arvola 2002; Miaskiewicz and Kozar 2011; Nielsen and Storgaard Hansen 2014), their wider proliferation is prohibited by the practical issues of time and cost of creation, as well as the concern of credibility of manual data collection and reliability concerns of qualitative analysis (Chapman and Milham 2006). In particular, the traditional way of creating Personas involves lengthy and costly manual work of focus groups and surveys. From our discussions with practitioners working in marketing agencies (An et al. 2017), we have learned that a thorough Persona creation process can take several months and cost between $80,000 and $120,000, when ordered from professional market research consultancies. Yet, these expensive Personas can quickly stale, if and when shifts in the underlying customer base take place. This phenomenon, known as concept drift in statistical computer science (Widmer and Kubat 1996), applies especially to cases where the underlying customer or user behavior is promptly changing, e.g. online content consumption (Thorson 2008).

Particularly, this is critical for those who do business decision-making based on Personas, but also for areas such as public health where content consumption over social media has an increasing role for information dissemination (Fernandez-Luque and Bau 2015). Therefore, an improved method of creating Personas is required for

collecting, processing, and presenting the real-time information to generate timely and accurate Persona profiles. Furthermore, content creators working for typical start-ups and small-and-medium sized enterprises (SMEs) cannot afford the sizeable investment of a full-scale Persona project costing up to tens of thousands of dollars. Therefore, they either leave the Persona approach unused or haphazardly produce Personas that are inferior to data-driven Persona representations. By development of a methodology for automatically generating Personas, it becomes possible to 'democratize Personas,' i.e. bring them in the reach of more decision makers in more areas in more organizations, thereby enhancing customer-oriented decision making and helping organizations better achieve their objectives. Overall, we postulate that better Personas result in higher user acceptance and traction from content creation teams and individuals working with online content in a professional manner. By content, we refer to videos, social media posts and web pages.

Current Drawbacks with Online User Analytics

Much of online user analytics is based on a collection of various user analytics numbers, often at a very low and simple level of counts and ratios but at great quantity. Understanding these counts and ratios, and then relating them to higher numerical analysis for the evaluation of key performance indicators (KPIs) (Jansen 2009) can be quite challenging, especially with large volumes of data. There is a plethora of tools (e.g., SimilarWeb, Google Analytics), methods (e.g., clustering, regression), services (e.g., comScore, HitWise) and software packages that one can employ. However, these tools, methods, services, and packages require a high degree of quantitative competence and analytical sophistication that the average professional end user rarely has. Additionally, these tools, methods, services, and packages often do little to reduce the complexity of numerical data in a manner that allows ease of use in decision making and communication. Dealing with numbers poses cognitive challenges for individuals who often do cannot recall many numbers at a time (Miller 1956). Personas, in contrast, are argued to be more intuitive and immersing forms of data representation (Cooper 2004; Matthews et al. 2012; Nielsen 2004).

Automatic Persona Generation as a Solution to Current Drawbacks

Interestingly, by combining Personas with user analytics data, the strengths of each approach can help address the shortcomings of the other. Personas are conceptually easy for people to understand, but they are slow to create and not be easily updated. User analytics data is current and easily collected, but it is cumbersome to employ and not easy to comprehend by analytically non-sophisticated end users. However, as illustrated in Fig. 8.1, the combination of the two leverages the strengths of both approaches, with limited degradation to the advantages of each.

Personas automatically generated from user analytics data have all the strength of standard Personas, although currently perhaps not as rich in detail, with many of the strengths of user analytics. Automatically generated Personas are current, along with being accurate, being based on real data, and possessing some detail of granularity, in that one can generate as many Personas as the underlying data indicates. While

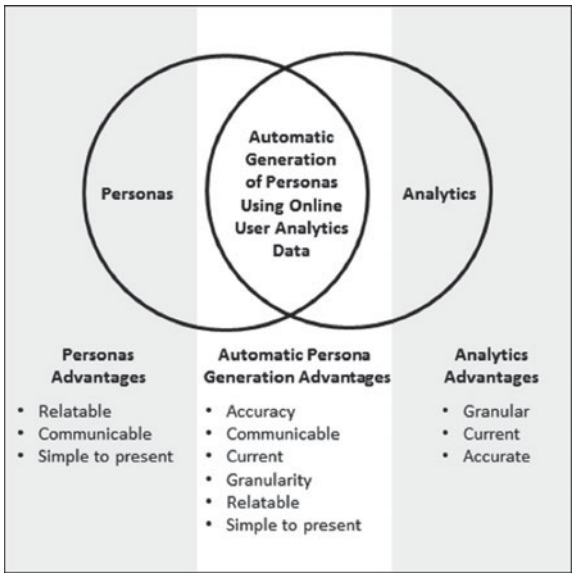

Fig. 8.1 Combining Personas with analytics numbers via automatic Persona generation leverages the strengths of both

automatically generated Personas are not expected to replace numbers in online content producers' decision making, they are aimed at providing intuitive descriptions of core users or customers and especially of the topical interests of these core groups. Numbers remain useful and, in fact, are asked for by the users (Salminen et al. 2018), and in our system, they are available both as raw data that can be downloaded by the end users and as contextual data breakdowns. The goal is to support the decision making of online content professionals that do prefer working with numbers, while giving easy access to underlying data also for number-oriented decision makers. In this manner, the system can provide value both for content producers and managers supervising content performance in various audience subgroups.

We define automatic Persona generation as a methodology for automatically creating Personas from online social media and analytics data (An et al. 2017), although one can expand the approach to other data sources as well, such as app stores, e-commerce customer base, and online gaming. Automatic Persona generation is specifically developed to address the aforementioned limitations of manual Persona creation and the employment of online user analytics data. Automatically generated Personas are (1) behaviorally accurate, as they are based on behavioral patterns inferred from users' interaction with online content and the demographic attributes of these users, (2) rapidly created once the behavioral data is collected from the online platforms, and (3) periodically updating through an automated loop of data collection and re-computation, based on near real-time data collection. For example, in a related study, we generated Personas from the YouTube channel of a large online media company with thousands of videos and hundreds of millions of views from an international audience consisting of viewers from 226 countries (Salminen et al.

2017). The creation of these Personas took literally only a few days—compared to traditional methods for Persona creation, it would not have been possible. In addition, (4) automatically generated Personas retain the privacy of individual users, both due to their nature as fictitious people and due to the use of aggregated data.[1] As such, automatic Persona generation has practical implications and is an exciting area of active, on-going research efforts.

8.3 APG: System and Methodology

8.3.1 Overview of System Functionality

By automatic Persona generation (APG), we refer both to the developed system that displays the generated Personas to end users in a client organization and to the methodology of generating these Personas. Thus, we will explain both in the following subsections. As mentioned earlier, we have made several fundamental steps in accomplishing the vision of generating fully automated Personas. We have thus far developed a stable and robust system[2] using open source technologies, including Flask Web framework[3] and PostgreSQL database.[4] The system is processing hundreds of millions of user interactions from online platforms, including YouTube, Facebook, Twitter, and Google Analytics (GA). The system is currently deployed in news organizations, such as Al Jazeera English, AJ + Arabic, and AJ + San Francisco, with clients in other verticals, including retailing and aviation, being added at a steady pace. We aim at expanding the user base to other domains, including public health professionals, e-commerce and online marketing managers, and other decision makers dealing with vast amounts of online data.

APG has many system functions from data collection to processing and enrichment, ending with providing the end users with an interface to interact with the Persona profiles (Jung et al. 2017). Figure 8.2 shows an overview of the system-level functions: (1) configuration, (2) collection, (3) Persona generation, and (4) interaction.

In the configuration stage, APG manages organizations and their users. Each organization needs at least one content source and associated API[5] information, such as the Channel ID for the analytics platform for YouTube, access token for Facebook, and so on. Depending on the organization's requirements, the collection period can be

[1] For example, instead knowing that a person named John Meyers watched a video about women's rights in Pakistan, we only know how many times the group he represents, e.g. [Male, 35–44, London] watched a particular video.

[2] A live demo is available at https://Persona.qcri.org/.

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

[4] https://www.postgresql.org/.

[5] Application programming interface, used for accessing a remote system; APG uses APIs for data retrieval.

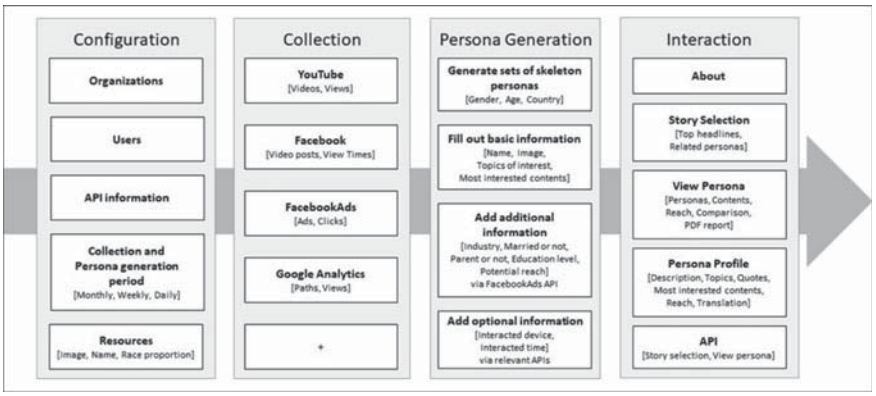

Fig. 8.2 From Persona creation to interactions with end users

adjustable to a monthly, weekly, or daily collection. In addition, APG stores resources such as image, name, and race proportion for new Personas. In the collection stage, once the configuration has been completed, the system collects the analytics content and interactions periodically. Currently, APG utilizes various content sources such as YouTube, Facebook, Facebook Ads and GA. Each content source has its own type of interactions. For instance, Facebook API and GA API enable us to harness view times of each video post and views of each page, respectively. The data is collected at an aggregated level, meaning the privacy of individual users is preserved.

After collecting the content and its interactions, APG generates a set of skeleton Personas. Each skeleton Persona has three demographic attributes: gender, age, and living country. Using this information, APG finds an appropriate name and image from a separate database (the names originate from country-level Census data, where available, and the copyrights for the pictures have been purchased from stock photo services). The system then retrieves topics of interest, calculated for each Persona based on the content they interact with (see the next section for more details), and examples of the content the Persona has shown the most interest in. Additionally, utilizing the demographic and topical information, APG retrieves further information, such as level of education and potential reach, via the Facebook Audience API. Finally, the system retrieves optional information such as device and time of interaction through relevant platform APIs, which completes the generation of Personas.

After generating Personas for an organization, APG enables the users (e.g., online content creators, producers or managers) of the organization to explore their own Personas in various ways. First, users can explore information about APG in the "About" page. Second, they can submit their own story on the "Submit Story" page in order to find matching Personas (this feature was developed for content creators using the system). Third, the "View Persona" page provides users with several ways to search for Personas: by the organization, data source, type, or number. For example, if they choose "Comparison" as a view type, they can compare temporal sets of Personas.

Moreover, each Persona can be explored in detail in the "Persona Profile" page. A report function provides a printable report of the Personas in portable document format (PDF).

8.3.2 Persona Generation Process

Overview of the Process

To generate Personas from aggregated social media data, our methodology undertakes the following steps:

1. Data collection via online analytics APIs (currently: YouTube Analytics, Facebook Insights, Google Analytics)
2. Data preparation by aligning content interaction patterns with sets of users
3. Identifying distinct content interaction patterns
4. Identifying impactful demographic groups from the set of distinct content interaction patterns
5. Creating skeletal Personas via demographic attributes from the data
6. Enriching the skeletal Personas with rich Personas description (e.g., name, picture).

APG Personas take into account both behavioral and demographic variation within the user base. To identify online content interaction patterns, we apply non-negative matrix factorization (NMF) (Lee and Seung 1999) in association interactions with the content. These interaction patterns are inferred from the user group and the content matrix then become the basis for Personas, to which we add user attributes, including topics of interest, demographics, photos, names, and other Personal details, in order to build the rich and authentic Persona profiles that are presented to the end users of our system.

Automatic Data Collection

The modern social media platforms provide access to user data via APIs. However, this data is strongly aggregated, i.e. bucketed into groups, to safeguard the privacy of individual users. An example of the bucket is [Female, 24–35, USA], describing the gender, age, and location of the users, rather than individual demographics of a specific user. The platforms gain this information from the users when they register, or by methods of statistical inference. Different interaction metrics are then shown for each group (e.g., clicks, views, or other available interaction metrics).

Table 8.2 shows an example of the data structure available at online analytics platforms.

The upper limit for the number of combinations of demographic groups is, therefore, [ages × genders × locations]. In the case of YouTube, for example, there are 4046 such combinations. Other online analytics platforms, such as Facebook and GA, follow a similar pattern in structuring their user data, as this way of data presentation

Table 8.2 Example of data structure used by APG

	Content 1	Content 2	Content 3	Content 4
Age A, Gender A, Location A	100	150	1000	200
Age A, Gender B, Location B	125	201	50	300
Age B, Gender B, Location A	200	500	33	145

Bolded cells indicate content a demographic group preferred (numbers are view counts)

is an industry norm. Via the APIs of the online platform, we can collect this bucketed data for different contents and different interaction metrics. For example, in the case of YouTube, we consider videos and their view counts, whereas, in GA, we consider pages and number of sessions to capture the visitor interactions with the website content as a whole. This detailed breakdown of user analytics data is accessible only to authorized users of the online analytics platforms. In practice, accessing the data requires collaboration with organizations that desire to have automatically generated Personas created from their audience statistics. The following section describes how we automatically analyze such data in order to build the skeletal Personas.

Data Processing and Persona Generation

Once the data has been collected from an online platform, we transform it into a matrix describing the interaction of users with the online content. We define $\mathbf{V}$ as the $\mathbf{g} \times \mathbf{c}$ matrix of $\mathbf{g}$ user groups and $\mathbf{c}$ the online content. The element of the matrix $\mathbf{V}$, V_{ij}, is any metric that reflects the interaction of the corresponding user group G_i for content C_j. For example, in the case of YouTube Analytics, V_{ij} is a view count for a particular video, C_j from user group G_i. The user groups are defined by gender, age range, and country (e.g., Male, 25–34, Finland). With $\mathbf{V}$ as the basis, we can infer a number of latent patterns by a form of decomposition (Lee and Seung 1999). These patterns constitute the basis of the Personas, as they capture the users' preferences for different content. The skeletal Personas are then enriched with additional information to produce a complete Persona profile.

In Fig. 8.3, the "Persona Profile" contains basic information of the Persona, including name, gender, age, and country. The "About Persona" includes a textual description of the Persona, generated based on a dynamic template. The "Topics of interest" describes the Persona's interests based on the topical classification of the content the Persona has viewed. The "Most Viewed Videos" lists the content the Persona has a distinct interest in. Finally, "Potential reach" is calculated by querying the Facebook Marketing API with the location, gender, age, and interests corresponding to the Persona, and reflects the total audience size of similar people.

Fig. 8.3 Example of an automatically generated Persona. Highlighted areas include [A] Image and demographic attributes, [B] Text description, [C] Topics of interest, [D] Descriptive quotes, [E] Most interesting content, and [F] Audience size

8.3.3 Applicability of APG in Varied Contexts

APG lends itself to many types of data. More specifically, the NMF approach is applicable across (1) data of different granularity (ranging from aggregate to individual level), (2) different content types (e.g., videos, web pages, advertisements, etc.), and (3) interaction metrics (e.g., views, clicks, downloads, purchases, etc.). In addition, APG can be applied to any dataset where the matrix **V**, which reflects user interaction with contents, can be defined. For example, if a mobile app store provides analytics data for app downloads by different user groups can be defined as the number of downloads from a particular user group for a given app. In such a case, APG can find Personas of that app marketplace.

Note that the user group can be replaced by an individual user if such data is available. In addition to online analytics/social media data, the NMF approach works with other types of data, including in-house data that is being used in many B2B environments, including law, medicine, engineers or other professions. The main requirement is a customer/user database with (a) a behavioral metric associated with each user/customer or group, and (b) demographic information, either at aggregated or individual level (associated with the behavioral metric). In conclusion, the core methodology can generalize across different sources of online analytics data, as well across different granularity of data and behavioral metrics, making it applicable to a variety of professional contexts. At the same time, we are noting that the APG Personas are currently best suited for online content professional due to their proven application on content interaction patterns. Future versions of APG could combine an organization's in-house data (e.g., customer relationship management database, CRM) with social media and online analytics data, resulting in even more useful, valuable and practically relevant Personas.

8.4 Setting the Research Agenda for APG

We have presented an overview of the current development of the APG system and methodology to demonstrate the feasibility of melding Personas and online user analytics data. From this base, we now present open research questions for further exploration. Figure 8.4 illustrates the research streams relating to APG, and they are explained thereafter.

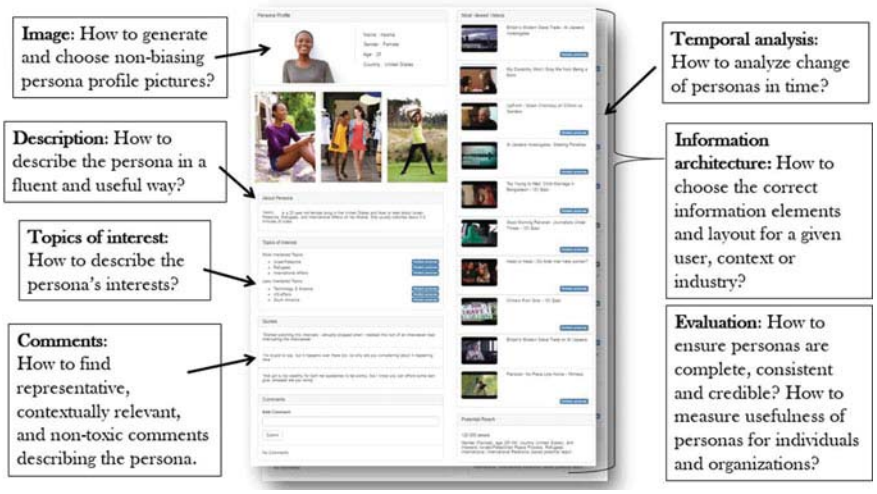

Fig. 8.4 Research agenda for automatic Persona generation

8.4.1 Design Philosophy of Automatically Generated Personas

A Persona profile typically includes a photo, name, description, and other information deemed relevant for the end users. Our automatically generated Personas are based on these basic information elements, including a picture, name, written description, topics of interest, and descriptive comments. We tailor the Persona profiles for clients. Given that our current client base is content creators, we provide two additional information elements: (1) most viewed YouTube videos, and (2) potential reach, obtained via Facebook Marketing API.[6] Our core tenet is that every information element (e.g., image, name, topics of interest, etc.) in the Persona profile, along with the whole profile representation itself, is subject to improvement by related research in computational techniques and automation.

The overall objective of the research project, from the computer and information science perspectives, is to discover better ways to computationally process and choose accurate and useful representations from vast amounts of online data. We embody this aspect in the motto *"Giving faces to data,"* which captures the objective of our research and development efforts. Table 8.3 clarifies the key purpose of improving each element of the Persona profile.

8.4.2 Choosing and Generating Profile Information

Automatic Image Generation

Currently, we purchase the Persona photos through professional photo selling websites. However, it is not always easy to find an appropriate photo for a particular demographic group, which is specifically gender, age, and ethnicity. In addition, there is a considerable cost factor, especially when expanding to cover more nations worldwide—the combinations of age, gender and location make the requirement for new photos excessive, counting up to thousands of photos. To save time and funds by dynamically generating suitable profile pictures is our motivation for exploring generative models for image creation. Developments in deep learning, particularly generative adversarial network (GAN) (Goodfellow et al. 2014) has been applied in modifying a given face to be older or younger. We are envisioning extending this line of work to support a wide range of visual attributes for generating photos that are appropriate for each Persona considering the age, gender, and ethnicity.

Selection of Name, Age and Ethnicity

Our current assignments of name and age are straightforward. We select the age from the age groupings of the social media and online analytics platforms. We select an age and gender appropriate name from an internal database which consists of

[6] https://developers.facebook.com/docs/marketing-apis.

Table 8.3 Rationale for improving each element of the automatically generated persona profile, and connections to research *Fields*

Persona element	Description	Goal or benefit	Field of research
Image	Using deep learning to generate Persona profile pictures	Generating artificial but realistic images while considering the underlying demographic variables (age, ethnicity, gender). Eliminates the need for manually downloading stock photos, thus resulting in time and cost savings	Computer vision, generative adversarial networks
Description	Describing the person in a fluent way with attributes relevant to decision makers	Using natural language processing and text generation to summarize numerical online data for the end users, thus increasing the appeal of the Persona descriptions	Natural language processing; text summarization and generation
Topics of interest	Creating topic classifications of online content and discovering probable interests across online platforms	Creating a scalable taxonomy, covering multiple domains and eliminating the need for creating organization- or industry-specific taxonomies each time a new one is added	Topic modeling, entity resolution, data mining and mapping
Quotes	Finding representative comments describing the Persona	The goal is to increase immersion to the Persona while preventing distraction from useful information	Social computing; automatic detection, classification and filtering of online hate
Story selection	Predicting and choosing content for Personas or content creators	Enables to formulate headlines correctly before publishing the stories; the goal is to enable online content producers to test content before publishing it	Prediction, recommendation engines, topic modeling

(continued)

Table 8.3 (continued)

Persona element	Description	Goal or benefit	Field of research
Temporal analysis	Observing change in Personas over time	Identifying trends in the underlying customer base, so that appropriate actions can be taken by end users when there are shifts in the content preferences of a Persona	Stream detection, concept drift, tensor factorization, time-series analysis
Information architecture	Choosing the correct information elements and layout for a given user or industry	Showing the right information to the decision maker in the right format, so that the system satisfies information needs of the organization and individual users in it, leading to improved decision making	Human-computer interaction, information design, usability, web analytics, user profiling
Evaluation	Validating accuracy and usefulness for individuals and organizations	Ensuring that Personas are reliable and credible, so that they can be trusted in real decision-making situations	Case studies, crowdsourcing; computational social science

typical names for people born in a certain country in a certain year. This database is manually compiled by searching online databases for Census data and similar data describing typical names in given periods. Certainly, there are many improved approaches that one can take for the selection of name, age, and ethnicity. In terms of ethnicity, the selection poses a significant challenge, as most platforms do not record this information. At present, we leverage the location (i.e., country) and infer the ethnicity probabilistically from Census data. However, there are alternative approaches to tackle the selection, such as analyzing profile images of real users or perform advanced textual analyses of social media comments (Nguyen et al. 2013).

Fluent Textual Description

The textual description included with a Persona profile, typically contains name, age, and gender, along with other demographical information, such as marital status, education level, career, etc. However, there have been surprisingly few studies into what information should be included and how this information can be utilized in a given context by a group of end users. This lack of prior work speaks to a need for HCI studies, such as eye-tracking or other bio-sensor type studies, along with ethnographic investigations of actual users of Personas in the workplace. In addition,

generative text algorithms could be applied in an attempt to automatically generate fluent descriptions of the Personas from data (Hu et al. 2017; Zhang et al. 2017).

Topics of Interest

Along with demographics, the advantage of using online data is that there is behavior analytics, including interactions with content. The overall goal is to develop an algorithm that can automatically and accurately classify the social media content based on content features, such as video titles, thumbnails, or descriptions, for example. Cross-platform data mapping becomes relevant in two cases: (1) when we would like to create Personas that describe user behavior along the same dimensions and metrics on different platforms; and (2) when different platforms have complementary information. For example, Platform A has information on demographics and topics, and Platform B has information on topics and Personality. Then, we need to find ways to reliably map these pieces of information if we want to include all of them in the Persona profile. We cannot reliably infer Personality traits from video consumption patterns, but the reliability of doing so is much higher with written text (Nguyen et al. 2013). To create rounded Personas, including attributes such as Personality and goals, we need to resort to using several platforms for the inference. Using only one platform at a time for Persona generation is a limitation of the current system (Jung et al. 2017).

Selecting and Filtering Quotes

Descriptive quotes are traditionally included in Persona description to illustrate the attitudes of the Persona (Cooper 2004). However, we observed during a user study that the interpretation of the Persona is highly influenced by the social media quotes (Salminen et al. 2018). For example, one of the participants noted: "I don't take her seriously because of the quotes". We also observed that the Persona can be perceived as something of a monster when we show unfiltered, hateful comments. In another case, a participant of the user study thought the Persona is shallow when there were quotes about nail polishing. To counter this sort of issues, we have developed three criteria for the quotes:

1. **Representative**: the selected comments accurately represent the Persona in question, primarily matching the behavioral patterns, topics of interest and other available identifiers
2. **Useful**: the selected comments increase customer insight and thus have the potential to be helpful for the end users of the system
3. **Non-distracting**: the selected comments are not purposefully offensive, so that the attention of end users of Personas remains focused on critical information (e.g., topics of interest).

In addition to filtering out the toxic comments (i.e., the non-offensiveness criteria), we should ensure the comments have a high probability of corresponding to the Persona (i.e., the relevance criteria). Prior work has shown that links from credible sources can enhance the perception of social media comments (Aigner et al.

Fig. 8.5 Example of quotes of a Persona in APG

2017), and thus, it would be interesting to investigate the interplay in the reverse direction in this domain. Thus, the question is: How do we determine, based on the social media comments, the demographics of the commentator and select only the comments that meet the Persona attributes? Moreover, in filtering out toxic comments, we should not resort to manipulating the data. In fact, if the data contains a large number of toxic dialogue, to maintain the integrity principle we should display those comments, regardless of them being offensive. This is the authenticity principle our Personas should follow. Evidently, the filtering problem represents a trade-off between authenticity and user experience. Figure 8.5 exemplifies our current quote system.

As a potential investigation into the trade-off mentioned above, we are interested in carrying out a user study where we would purposefully show Personas with toxic comments to end users and let them describe the Persona, in comparison to a filtered view where the comments are cleaned. We already know the comments influence the interpretation of the Persona (Salminen et al. 2018) and prior work has shown that social signals impact aspects of information use (Badache and Boughanem 2014), but we are interested in fortifying that argument and exploring the topic further that can lead to interesting findings. Finding impact not only Persona analytics but also regarding polarization, as we could evaluate how the counterparts of an argument are interpreted with or without hatefulness. The latter condition could result in avoidance of many unnecessary feuds, and if confirmed by such a user study, it would suggest that strong moderation is needed for online social platforms.

8.4.3 Deciding Information Architecture

Regarding what information to include and how to present it, there are two issues we have identified.

Selection Problem

This problem is defined by choosing which information to include in the Persona profile, i.e., the structure of the layout. There are more possibilities for information elements that can be shown due to cognitive and screen space limitations. Three approaches can potentially address this issue: (1) first, defining common information needs for a given industry or domain, and choosing information elements accordingly. For example, information needs in e-commerce typically differ from the type of information that journalists would expect from Personas; e.g., consumers' search keyword information is important for search-engine advertisers (Jansen and Mullen 2008; Jansen and Spink 2006), but not so relevant for news professionals. Another option is (2) Personalizing the information based on *individual* user needs, i.e., identifying the end user's usage patterns or preferences and changing the information elements to automatically match these information needs (a type of website morphing (Hauser et al. 2009). Third, we could apply (3) self-selection, in which the end users can build their own Personas by choosing their preferred information elements from all the available information elements. Further research is required to match the information supplied with the information asked for by the end users, and also to clarify to which degree there is a commonality and, on the other hand, idiosyncrasy across professional verticals.

Aggregation Problem

This problem deals with choosing which data—i.e., the content of the information elements—to include. Selection of any information easily leads to biased interpretations, as we found in a user study in which the chosen Persona images and quotes greatly influenced the users' perception of the Persona (Salminen et al. 2018). We propose two potential solutions: (1) removing ambiguous and controversial informational elements to de-bias the end users (e.g., not showing contextual photos), and (2) purposefully introducing diversity to display the variation in the underlying user base. As demonstrated in Fig. 8.6, over-aggregation can be solved by introducing an additional layer of depth in the information hierarchy. Such an approach could be used to mirror each active information element in a "deeper layer" that holds breakdown information. By showing such contextual information, we may be able to curb the tendency of Personas to evoke stereotypical thinking—however, confirming this requires further research. An example in the APG Personas is the demographic information: we choose the Persona's age and gender, but, in fact, there is an underlying *distribution* of all ages and genders. For example, "*Samantha, 25 from New York*" can have a degree of topical similarity to a 42-year-old man from Doha, for example, but the algorithm is forced to choose the most impactful group due to the fact that Personas can only have one visible combination of [age, gender, location]. In this

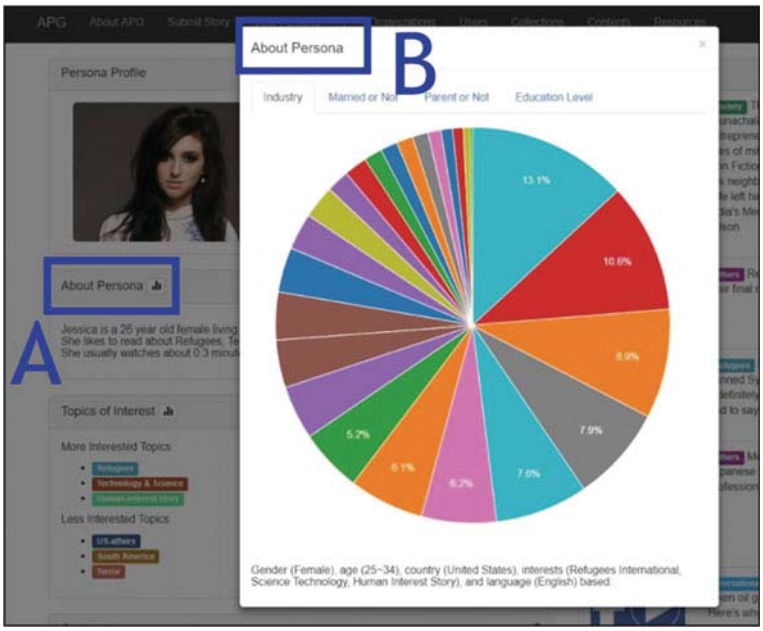

Fig. 8.6 Example of showing diversity to solve the aggregation problem of automatically generated Personas. **a** In the picture shows the high-level information, and clicking the details icon will reveal a break-down of information, **b** about the user or customer

case, we could show a breakdown of all the age, gender, location combinations as a separate view.

The drawback of the "forced diversity" approach is that it may reduce the sense of empathy toward the Persona, as it becomes more obvious to the end users that the Persona actually consists of several sub-groups. It is, however, unclear whether this takes place or not, and further research is needed to clarify the matter. In the worst scenario, introducing variety results in the loss of coherence, so that the Persona becomes a fragmented group of individuals instead of one person to relate to (Nielsen 2002). To maintain the immersing benefits, consistency and coherence are, according to our view, necessities. In sum, the power of the Persona system in framing the interpretation of the end users seems consistent with other works discussing the effects of automated systems (e.g., recommendation engines) on users' interpretations and choices vis-'a-vis the power of algorithms, prompting for further research.

Relating to defining the information elements included in the Persona profile, it seems that filtering should be done based on user modeling, user studies, or field studies in specific domains. Moreover, after defining the desired information content, there is a question of its retrieval. As mentioned earlier, our Personas consist of basic information elements, such as image, name, and topics of interest. However, social media platforms, such as Twitter, enable us to infer other attributes (Schwartz et al.

2013; Zagheni et al. 2014). It has been shown possible to infer, for example, a user's political affiliation, Personality traits, socioeconomic status, relationship status, with varying degrees of accuracy. The applied methodologies are wide, including graph analysis, matrix factorization, natural language processing, and so on. Feature selection and extraction seem critical here, requiring a level of subject matter knowledge. Moreover, cross-platform data mapping, e.g. in the form of Entity Resolution problem (Feldman 2013), is also needed to combine social media and online analytics datasets with satisfactory validity.

8.5 Summary of Challenges in Automatic Persona Generation

This chapter discussed the computational problems identified with automatic Persona generation, a novel way of creating accurate and meaningful Personas from online analytics data. Our objective was (1) to describe the overall research project, (2) to identify sub-problems relating to achievement of the overall goal of generating accurate and meaningful Persona profiles for end users of a given organization, and (3) to outline some potential research approaches to solving them. Overall, our conclusions are the following.

Research on automatic Persona generation combines various fields ranging from algorithmic, "hard-core" computer science to softer user studies with an interpretative touch. As such, APG is an excellent platform for mixed-method research. Also, as Personas are applicable across industries, we are actively seeking to expand the method's application to new fields, e.g., e-health, airline industry, e-commerce, and marketing. Our vision is to achieve completely automatic Persona generation from several online platforms via automatic data collection and processing on a robust infrastructure that scales with an increasing number of end users.

In our effort of outlining the research agenda, we have also defined several criteria and principles for the development of automated Personas. They can be viewed as "simple guiding principles" (Bürgi et al. 2004) for system development. Just as Personas are to be deployed as useful mental models for practical decision making, these user-oriented principles should guide our research and development efforts. They are as follows:

1. **Consistency**: the independent elements of the Persona profile are matching one another to create a coherent, logical and truthful portrayal of the underlying user segment.
2. **Relevance criterion**: the shown information elements as a whole and also the pieces of information within each element are chosen to be of immediate use and relevance to the end users of the system.
3. **Non-offensiveness**: whenever possible, we prevent the psychologically harmful information elements from showing to end users; however, while doing so, we cannot deviate from the authenticity principle.

4. **Authenticity**: we do not knowingly manipulate the Persona representations to deviate from the data, even if this means showing information that is offensive to the end users.
5. **Context**: we are showing a contextual layer of information of the Persona to the end users, thereby making clear it is based on a group of people (a distribution).

We remind that the current work is part of an on-going, continuously developing a research project, in which novel sub-problems are identified in interaction with actual users of the system. While this introduces certain unpredictability to our research and development agenda, it also provides one worthy advantage, namely avoiding the hazardous 'closed loop problem' in which our own preferences, rather than user feedback, would dictate development priorities (Salminen 2014). As such, we are looking forward to developing the system, and the associated research agenda, in close connection to end users according to their domain-specific information needs.

8.6 Conclusion

We postulate that automatically generating Personas from actual user analytics data is a fruitful and impactful research area with potential for both focused disciplinary and also cross-disciplinary research, ranging from algorithmic to HCI to analytics to information to cognition. The research agenda outlined in this perspective manuscript addresses existing issues in two senses.

First, while often rich in detail, Personas created via qualitative methods are slow, expensive, and difficult to update without a complete replication of the creation process. Automatically generating Personas from online data can address these current shortcomings, being fast, inexpensive, and easily updated, while preserving the privacy of individual users. Second, online user analytics data, while detailed and readily available on many online social media and other platforms, is cumbersome and unwieldy to use for analytically unsophisticated users. Automatically generating Personas from online data can address these current shortcomings by simplifying the presentation while still enabling access to the data itself, making the APG system useful for professionals working with online content creation.

Finally, regarding the open research questions explicated in this work, our key interest includes improving the system for end users, while at the same time engaging in interesting and productive research for improving the quality of the automatically generated Persona profiles. Data-driven Personas are also opening new research avenues for social science experiments in computational social science (social computing). For example, through Personas, we can examine the innate biases of end users of customer analytics. Studies have shown that Personas open a peripheral route into individuals' thinking, particularly revealing end users' subjective perceptions about the audience or particular user groups (Hill et al. 2017; Salminen et al. 2018). By creating controlled Persona variations from the underlying user data, we

can conduct social science studies that examine how content creators perceive and respond to Personas of different types. For example, comparing "toxic" and neutral Personas is among our future research ideas in this space.

Acknowledgements We would like to thank the employees of the Al Jazeera Media Network, Qatar Airways, and Qatar Foundation who have collaborated with us on this project.

References

Agarwal R, Dhar V (2014) Editorial—big data, data science, and analytics: the opportunity and challenge for is research. Inf Syst Res 25(3):443–448

Aigner J, Durchardt A, Kersting T, Kattenbeck M, Elsweiler D (2017) Manipulating the perception of credibility in refugee related social media posts. In: Proceedings of the 2017 conference on conference human information interaction and retrieval. ACM, New York, NY, USA, pp 297–300

An J, Haewoon K, Jansen BJ (2016a). Towards Automatic Persona Generation Using Social Media. In *Proc. of The Third International Symposium on Social Networks Analysis, Management and Security (SNAMS 2016)*, The 4th International Conference on Future Internet of Things and Cloud. 22–24 August

An J, Kwak H, Jansen BJ (2016b) Validating social media data for automatic Persona generation. In: Proceedings of the second international workshop on online social networks technologies (OSNT-2016), 13th ACS/IEEE international conference on computer systems and applications AICCSA 2016, 29 Nov–2 Dec

An J, Haewoon K, Jansen BJ (2017) Personas for content creators via decomposed aggregate audience statistics. In: Proceedings of Advances in Social Network Analysis and Mining (ASONAM 2017), 31 July

Badache I, Boughanem M (2014) Harnessing social signals to enhance a search. In: 2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT) (Presented at the 2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT), vol 1, pp 303–309

Blomquist, AAsa, Arvola M (2002). Personas in action: ethnography in an interaction design team. In: Proceedings of the second Nordic conference on human-computer interaction, pp 197–200

Bürgi P, Victor B, Lentz J (2004) Modeling how their business really works prepares managers for sudden change. Strat Leadersh 32(2):28–35

Chapman CN, Milham RP (2006) The Personas' new clothes: methodological and practical arguments against a popular method. Proc Hum Factors Ergon Soc Annu Meet 50(5):634–636

Chapman CN, Love E, Milham RP, ElRif P, Alford JL (2008) Quantitative evaluation of Personas as information. Proc Hum Factors Ergon Soc Annu Meet 52(16):1107–1111

Cooper A (2004) The inmates are running the asylum: why high tech products drive us crazy and how to restore the sanity, 1st edn. Sams—Pearson Education, Indianapolis, IN

Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM 56(4):82–89

Fernandez-Luque L, Bau T (2015) Health and social media: perfect storm of information. Healthc Inform Res 21(2):67–73

Friess E (2012) Personas and decision making in the design process: an ethnographic case study. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI'12). ACM, New York, NY, USA, pp 1209–1218

Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manage 35(2):137–144

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S et al (2014) Generative adversarial networks. arXiv:1406.2661 [cs, stat]. http://arxiv.org/abs/1406.2661. Accessed 27 Feb 2018

Goodwin, K. (2011). Designing for the digital age: how to create human-centered products and services. Wiley, New York

Guo G, Zhu F, Chen E, Liu Q, Wu L, Guan C (2016) From footprint to evidence: an exploratory study of mining social data for credit scoring. ACM Trans Web 10(4):1–38

Hauser JR, Urban GL, Liberali G, Braun M (2009) Website morphing. Market Sci 28(2):202–223

Hill CG, Haag M, Oleson A, Mendez C, Marsden N, Sarma A, Burnett M (2017) Gender-inclusiveness Personas vs. stereotyping: can we have it both ways? In: Proceedings of CHI '17, ACM Press, pp 6658–6671

Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing EP (2017) Controllable text generation. ArXiv preprint arXiv:1703.00955

Jansen BJ (2009) Understanding user-web interactions via web analytics. Synth Lect Inf Concepts Retrieval Serv 1(1):1–102

Jansen BJ, Mullen T (2008) Sponsored search: an overview of the concept, history, and technology. Int J Electron Bus 6(2):114–131

Jansen BJ, Spink A (2006) How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Inf Process Manage 42(1):248–263

Jansen BJ, Sobel K, Cook G (2011) Classifying ecommerce information sharing behaviour by youths on social networking sites. J Inf Sci 37(2):120–136

Jansen BJ, An J, Kwak H, Salminen J, Jung S-G (2017) Viewed by too many or viewed too little: using information dissemination for audience segmentation (pp 189–196). In: Presented at the association for information science and technology annual meeting 2017 (ASIST2017), Washington DC, USA

Jenkinson A (1994) Beyond segmentation. J Target Measure Anal Market 3(1):60–72

Jung S-G, An J, Kwak H, Ahmad M, Nielsen L, Jansen BJ (2017) Persona generation from aggregated social media data. In: Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems (pp 1748–1755). ACM, New York, NY, USA

Kwak H, An J, Jansen BJ (2017) Automatic generation of Personas using youtube social media data (pp 833–842). In: Proceedings of the Hawaii international conference on system sciences (HICSS-50). 4–7 Jan, Waikoloa, Hawaii

Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791

LeRouge C, Ma J, Sneha S, Tolle K (2013) User profiles and Personas in the design and development of consumer health technologies. Int J Med Inform 82(11):251–268

Matthews T, Judge T, Whittaker S (2012) 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, New York, NY, USA, pp 1219–1228

McGinn JJ, Kotamraju N (2008) Data-driven Persona development. In Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 1521–1524

Miaskiewicz T, Kozar KA (2011) Personas and user-centered design: How can Personas benefit product design processes? Des Stud 32(5):417–430

Miaskiewicz T, Sumner T, Kozar KA (2008) A latent semantic analysis methodology for the identification and creation of Personas. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 1501–1510

Miller GA (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev 63(2):81–97

Nguyen D.-P., Gravel R, Trieschnigg RB, Meder T (2013) "How old do you think I am?" A study of language and age in Twitter. In: Proceedings of the seventh international AAAI conference on weblogs and social media (ICWSM). Cambridge, Massachusetts, USA

Nielsen L (2002) From user to character: an investigation into user-descriptions in scenarios. In: Proceedings of the 4th conference on designing interactive systems: processes, practices, methods, and techniques. ACM, New York, NY, USA, pp 99–104

Nielsen L (2004) Engaging Personas and narrative scenarios (vol 17). Samfundslitteratur. http://Personas.dk/wp-content/samlet-udgave-til-load.pdf

Nielsen L, Storgaard Hansen K (2014) Personas is applicable: a study on the use of Personas in Denmark. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 1665–1674

Nielsen L, Jung S-G, An J, Salminen J, Kwak H, Jansen BJ (2017) Who are your users?: comparing media professionals' preconception of users to data-driven Personas. In: Proceedings of the 29th Australian conference on computer-human interaction. ACM, New York, NY, USA, pp 602–606

Oviatt S (2006) Human-centered design meets cognitive load theory: designing interfaces that help people think. In: Proceedings of the 14th ACM international conference on Multimedia. ACM, pp 871–880

Pruitt J, Grudin J (2003) Personas: practice and theory. In: Proceedings of the 2003 conference on designing for user experiences. ACM, New York, NY, USA, pp 1–15

Rönkkö K, Hellman M, Kilander B, Dittrich Y (2004) Personas is not applicable: local remedies interpreted in a wider context. In: Proceedings of the eighth conference on participatory design: artful integration: interweaving media, materials and practices-volume 1 (PDC 04). vol. 1, ACM, New York, NY, USA, pp 112–120

Rönkkö K (2005) 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-volume 08 (HICSS '05), vol. 8. IEEE Computer Society, Washington, DC, USA, p 220

Salminen J (2014) Startup dilemmas—Strategic problems of early-stage platforms on the internet (Doctoral dissertation). Turku School of Economics, Turku. Retrieved from http://www.doria.fi/handle/10024/99349

Salminen J, Milenković M, Jansen BJ (2017a) Problems of data science in organizations: an explorative qualitative analysis of business professionals' concerns. In: Proceedings of International Conference on Electronic Business (ICEB 2017). Dubai

Salminen J, Şengün S, Haewoon K, Jansen BJ, An J, Jung S et al (2017b) Generating cultural Personas from social data: a perspective of middle eastern users. In: Proceedings of the fourth international symposium on social networks analysis, management and security (SNAMS-2017), Prague, Czech Republic. Accessed 26 Aug 2017

Salminen J, Kwak H, Santos JM, Jung S-G, An J, Jansen BJ (2018a) Persona perception scale: developing and validating an instrument for human-like representations of data. In: CHI'18 extended abstracts: CHI conference on human factors in computing systems extended abstracts proceedings, Montréal, Canada

Salminen J, Nielsen L, Jung S-G, An J, Kwak H, Jansen BJ (2018b) "Is more better?": impact of multiple photos on perception of Persona profiles. In: Proceedings of ACM CHI conference on human factors in computing systems (CHI'18), Montréal, Canada

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M et al (2013) Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS ONE 8(9):e73791

Scott DM (2007) The new rules of marketing. Wiley, Hoboken, New Jersey

Stauss B, Heinonen K, Strandvik T, Mickelsson K-J, Edvardsson B, Sundström E, Andersson P (2010) A customer-dominant logic of service. J Serv Manage 21(4):531–548

Thorson E (2008) Changing patterns of news consumption and participation: News recommendation engines. Inf Commun Soc 11(4):473–489

Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. Science 185(4157):1124–1131

Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. Mach Learn 23(1):69–101

Zagheni E, Garimella VRK, Weber I, State B (2014) Inferring international and internal migration patterns from twitter data. In: Proceedings of the 23rd international conference on World Wide Web, ACM, New York, NY, USA, pp 439–444

Zhang X, Brown H-F, Shankar A (2016) Data-driven Personas: constructing archetypal users with Clickstreams and user telemetry. In: Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 5350–5359). ACM, New York, NY, USA. Accessed 4 Nov 2017
Zhang Y, Gan Z, Fan K, Chen Z, Henao R, Shen D, Carin L (2017) Adversarial feature matching for text generation. ArXiv preprint arXiv:1706.03850