Social Network Analysis and Mining (2018) 8:54
https://doi.org/10.1007/s13278-018-0531-0

ORIGINAL ARTICLE

Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data

Jisun An[1] · Haewoon Kwak[1] · Soon-gyo Jung[1] · Joni Salminen[1,2] · Bernard J. Jansen[1]

Received: 21 December 2017 / Revised: 28 April 2018 / Accepted: 5 August 2018

Abstract
We propose a novel approach for isolating customer segments using online customer data for products that are distributed via online social media platforms. We use non-negative matrix factorization to first identify behavioral customer segments and then to identify demographic customer segments. We employ a methodology for linking the two segments to present integrated and holistic customer segments, also known as personas. Behavioral segments are generated from customer interactions with online content. Demographic segments are generated using the gender, age, and location of these customers. In addition to evaluating our approach, we demonstrate its practicality via a system leveraging these customer segments to automatically generate personas, which are fictional but accurate representations of each integrated behavioral and demographic segment. Results show that this approach can accurately identify both behavioral and demographical customer segments using actual online customer data from which we can generate personas representing real groups of people.

Keywords Web analytics · Social computing · Personas · Marketing · System design · Customer segmentation

1 Introduction

One use of social media and other web analytics data is customer segmentation (Jansen 2009), which is an approach for separating an overall customer population based on segment differences defined by a specific set of attributes. Customer segmentation is a common practice across many industries with the set of attributes utilized being relevant to the particular domain. Examples of such domains include marketing, advertising, education, and system design. E-commerce companies and other organizations rely on customer segmentation to target specific customer groups with content and products that the consumers within a segment would likely find relevant. Additionally, customer segmentation might also lead to a deeper understanding of customer preferences, needs, and wants by isolating what each segment finds most valuable. Based on these insights, organizations can more effectively engage with their customers, audience, or users. In software design, marketing planning, and advertising development, there are continuing efforts for identifying and assessing segments of people (i.e., customers, audience, or markets) to optimize some performance metric (e.g., the speed of task, buying preferences, or ease of use).

Major online social media platforms used for distributing content and other products present unique challenges for customer segmentation efforts attempting to rely on online customer data. The customer segmentation approach relies on identifying key attributes from which one can separate customers into segments (Cooil et al. 2008). Targeting customers via behavioral segmentation involves dividing the customer base based on their collective behavior. A behavior can be a single attribute

✉ Jisun An
jisun.an@acm.org

Haewoon Kwak
haewoon@acm.org

Soon-gyo Jung
sjung@hbku.edu.qa

Joni Salminen
jsalminen@hbku.edu.qa

Bernard J. Jansen
bjansen@hbku.edu.qa

[1] Qatar Computing Research Institute, HBKU, Ar Rayyān, Qatar

[2] Turku School of Economics, Turku, Finland

Published online: 23 August 2018

(e.g., viewing online content) or a set of behaviors (e.g., viewing online content, length of video watched, etc.), but it is typically focused on the way the segment responds to, uses, or engages with a product. Targeting customers via demographic segmentation includes segmenting the customers based on one or more differentiating characteristic that often includes, but is not limited to, gender, age, race, location, education, income, or career. However, most prior work in customer segmentation has focused on using individual website data, such as that available from Google Analytics, yet, there is an increasing effort to employ customer segmentation using social media data from the major online platforms. This data presents unique challenges, as it is typically aggregated to preserve the privacy of individuals, so methods must be employed to deal with the issues this aggregation causes in inferring customer attributes.

Our aim is to automatically create customer segments using real customer data aggregated by online platforms, meaning it is grouped according to customer attributes (e.g., Male, 25–34, South Korea) (Jung et al. 2017). This grouping is more common due to privacy concerns that prompt online platforms to provide only aggregated customer data, rather than session- or customer-level data. Aggregation complicates the customer segmentation generation process. In the research reported here, we investigate using aggregated social media data for isolating customer segments based on both the behaviors and on the demographics of those customers and then linking the two customer segments groupings for a complete representation of the customer base. Therefore, we propose, develop, implement, and evaluate an approach for mining privacy-preserving aggregated statistics of the customer base, which we can use to generate data-driven customer segments that are easily interpretable by web information analysts. To demonstrate the impact and applicability of this customer segmentation research, we leverage that information to develop a system to automatically generate personas for the identified customer segments.

One advantage of this approach is that it can easily be adapted for a diverse range of organizations that create content for major online platforms since such aggregated customer statistics are the de facto standard provided by analytics interfaces of online platforms, such as Facebook Insights and YouTube Analytics. Our research differs from prior work in that earlier work in identifying customer segments (Jansen et al. 2011) or for the internal application of a private company (Zhang et al. 2016) had individual-level data. However, due to privacy and other concerns, online data from the major platforms is typically not individualized. Instead, the data has been already aggregated, typically along coarse attributes, such as gender, complicating the customer segmentation process.

2 Literature review

2.1 Customer segmentation

The concept of customer segmentation is attributed to Smith (1956), where the researcher advocated employing market segmentation along with the product segmentation that was common at the time. Since then, customer segmentation has been an ongoing research area (Bonoma and Shapiro 1984), as the availability of online data has greatly increased. Overall, customer segments arise from attributes that unify customers to form groups or separate some customers from others (Jenkinson 1994). There have been a variety of data and methods employed to create customer segments (Firat and Shultz 1997; Marcus 1998; Shapiro and Bonoma 1984). Given the availability of ample survey of literature articles in the area (Beane and Ennis 1987; Chéron and Kleinschmidt 1985; Foedermayr and Diamantopoulos 2008), we do not present a comprehensive review here but provide insights in the activity and variety of research in the customer segmentation area.

Website data has been used to segment customers into various revenue groupings (Ortiz-Cordova and Jansen 2012); this is an example of behavioral segmentation. Search query data has been used to classify the gender of searchers and then relate this demographic attribute to revenue generation (Jansen et al. 2013). Increasingly, customer segmentation processes are leveraging social media data for both behavioral and demographic grouping (Jansen et al. 2011) of customers. Tuna et al. (2016) examine the identification of segments from social media, specifically from customer attributes such as gender and age, among others. Dursun and Caber (2016) employ RFM (recency, frequency, monetary) analysis on data from a major hotel chain's customer relationship management system with results showing eight customer segments with the majority of the customers as 'Lost Customers,' staying for shorter periods and spending less relative to other segments. RFM analysis is a marketing technique used to quantitatively determine which customers are the best ones by examining how recently a customer has purchased (recency), how often they purchase (frequency), and how much the customer spends (monetary). Antoniou (2017) uses segmentation from online platforms in cultural heritage applications by extracting user personality and cognitive style profiles. Concerning the specific use of social media data, Kamboj, Kumar, and Rahman (2017) find that social use, hedonic use, and cognitive use positively influence the financial and market performance of firms.

2.2 Personas

As the motivation for our customer segmentation approach, we use it to develop a working system for automatic persona generation. Introduced to the design domain by Cooper

(2004) with follow-up refinement by Pruitt and Adlin (2006), a persona is a representation of an actual segment of customers presented as an imaginary person. The end product is a persona profile pertaining to the customer segment that the fictionalized person represents. Personas have expressed benefits beyond what numbers by themselves can provide for identifying customer segments (Pruitt and Grudin 2003). For several years, personas have been used in system development (Cooper 2004; Pruitt and Adlin 2005), product design (Goodwin and Cooper 2009; Smith 1956), and marketing (Revella 2015; Stern 1994), among many other fields and industry verticals. Personas are a continuance of efforts from a variety of domains for identifying, constructing, and assessing segments of people (i.e., customers, audience, etc.) to optimize some performance metrics (e.g., advertising engagement, the speed of a task, ease of use, the effectiveness of effort, sales, etc.). Personas are reportedly a part of design processes and industry workflows (Dharwada et al. 2007; Eriksson et al. 2013; Friess 2012; Judge et al. 2012; Nielsen and Hansen 2014) for both long- and short-term projects (Judge et al. 2012) with a reported positive return on investment (Drego and Dorsey 2010).

It is suggested that one develops personas from real data that is derived from actual people (Pruitt and Adlin 2006). Using actual customer data is crucial to making personas believable (Judge et al. 2012) and for designers to appropriately leverage personas. However, a recognized problem is that creating personas is always not a cheap or quick procedure, as the creation has historically involved ethnographic methods. As one-time data collection actions, the personas created can be quickly outdated without new rounds of data collection. Without real-time data, designers have no validation whether the personas are representative of current customers (Chapman and Milham 2006). These restrictions are especially acute in the situation of creating digital content for distribution via major online platforms (e.g., Facebook, Twitter, YouTube, etc.). The research that has been conducted in converting actual online customer data into personas is limited and is also quite sparse in actually creating personas from this data. In the marketing and advertising area, there is work on using large pools of online consumer data to segment markets (Clarke 2015); the work focused on overall design approaches but did not take the research to creating personas. For example, Jansen et al. (2011) used the data from nearly 35,000 customers on a social media platform to cluster customers based on how they share commercial information. However, the researchers did not use these results to generate personas, stopping at the segmentation level; although, they did assign descriptive names to each. In another work, Zhang et al. (2016) analyzed customer-level clickstreams to identify ten common workflows using hierarchical clustering. They then present five customer facets based on the probability of platform use that they then gave a name to. While this work is close to our work, the customer-level clickstreams are often not available to those online content creators, especially when using aggregated data from online platforms.

2.3 Synthesis of prior work

From a review of customer segmentation literature, research using customer-level segmentation data, especially to generate personas, is needed. With the potential customers in the millions or billions, traditional ethnography and related methods may not scale well and can be cost-prohibitive. While there have been some online data-driven approaches (Chiang et al. 2015; Zhang et al. 2016), they have used fine-grain customer-level data that is not often available and that potentially has privacy issues. Such approaches using individual-level data are not suited for most content creators who only see aggregated statistics via a platform's analytic tools.

Therefore, there are many unanswered questions concerning using social media analytics for customer segmentation and whether the findings from this segmenting process can be put to practical use. Can one isolate customer segments based on behavioral interaction on social media platforms? Can online data deliver demographic insights for customer segmentation? Can the customer segments be identified in real time? Can the customer segments be frequently updated? These are the questions that motivate our research.

Thus, this investigation continues a stream of research in generating customer segments and personas (An et al. 2016a, b, 2017; Jansen et al. 2017a; Jung et al. 2017; Kwak et al. 2017) from publicly available social media data, such as from Facebook (Jansen et al. 2016; Zhang et al. 2016) or YouTube (Jansen et al. 2017b) in which the approach was clustering and unsuccessful. Specifically, the research reported in this manuscript is an expansion of a four-page conference article (An et al. 2017). In this manuscript, we focus on the customer segmentation aspects of the research, expand the data sets employed in the analysis, increase the methods of evaluation, and showcase the application of our customer segmentation approach in the development of a system to automatically generate personas from large-scale, aggregated social media data. Personas generated from these methods can be used as-is or can be used in conjunction with data collected from more traditional persona creation methods, and they can be enriched further using qualitative methods (Salminen et al. 2017).

3 Research objectives

Our premise is that aggregated behavioral and demographic customer data, as well as privacy-preserving concerning consumers of a product, service, system, or content can be

collected from major online platforms and can be rapidly analyzed to identify customer segments that are usable for a variety of commercial purposes. From this premise, our goal is to develop a methodology to (a) mine aggregated large-scale privacy-preserving aggregated online customer data from major online platforms; (b) use this online data to identify distinct and impactful customer segments; and (c) to automatically generate personas with realistic descriptions and attributes that represent these key customer segments. We see several advantages to this approach, with this approach sufficing as either a standalone method for personas generation or in conjunction with conventional offline methods of persona creation. Therefore, our research objectives are:

1. Recognize discrete customer segments based on behavioral interactions with online content posted on major online social media platforms.
2. Identify the discrete demographic customer segments associated with each of these behavioral customer segments.
3. Integrate the associated behavioral customer segments and the demographic customer segments.
4. Demonstrate the practicality of this approach via a system to automatically generate personas representative of these customer segments.

For investigating these research questions, we rely on non-negative matrix factorization (NMF). The foundation and the encoding depend on what decomposition technique is employed. There are three matrix decomposition methods that are commonly used for the purpose presented here, namely principal component analysis (PCA), vector quantization (VQ), and non-negative matrix factorization (NMF) (Lee and Seung 1999). Concerning the actual technique, VQ, PCA, and NMF bring different decomposition outcomes by having different constraints in **W** and **H**. With VQ, each column in **H** has to be a unary vector. Therefore, only one entry in a given column in **H** has a non-zero value, and all others have to be zero (Gray 1984). This one-entry constraint makes **H** too simplified to explain meaningful behavioral patterns by a combination of content interactions. Consequently, VQ is not appropriate for our purpose. With PCA, the rows in **H** have to be orthogonal, and columns in **W** have to be orthonormal (Jolliffe 2002). PCA approaches an entry in **V** as a linear combination of the corresponding row and the column in **W** and **H**, respectively. Nevertheless, PCA entries in **W** and **H** can be either positive or negative. These positive and negative coefficients result in complex cancellations, making the results difficult to interpret, so PCA is inappropriate for our purpose. In contrast, NMF does not allow negative entries in **W** and **H**. As no subtraction led by the negative coefficients is permissible, we consider a linear combination as only an additional combination of bases. This non-zero restriction makes interpretation of the matrix decomposition straightforward; therefore, we choose NMF to extract shared content consumption patterns from the aggregated customer interaction statistics.

That being said, this research is novel in several respects. It is one of the first research efforts using online aggregated social media data at scale for customer segmentation. The data from such platforms is aggregated, unlike the data from prior work in identifying segments using individual-level data. Due to privacy and other concerns, customer data from the major online social media platforms is not individualized, as it is aggregated along typically coarse attributes such as gender, which complicates generating customer segments. Therefore, one must develop techniques to decompose this aggregated data for customer segment generation while still respecting data privacy. Our method is flexible in terms of the number of possible customer segments generated. Typically, customer segmentation and persona creation focus on a small number of segments or personas. While appropriate in certain environments, it may not be appropriate for organizations distributing products via major online platforms with millions or billions of worldwide customers. As our data sets are typically large, in the tens of millions if not more, we can validate our customer segments using quantitative methods. Also, our use of automatically generating personas from these customer segments is novel. Beyond our limited previous work (An et al. 2016a, b; Jansen et al. 2017b), we could locate no prior research pertinent to generating fully developed personas using aggregated data for those who distribute their products via major online platforms. Finally, the methodology presented here can be generalized to any organization that distributes content via major platforms. Therefore, the impact of this research is broad, and it is applicable to many domains.

4 Framework to identify customer segments

To develop customer segments from aggregated customer statistics, we first formulate the problem and clarify the setting, particularly the characteristics of the required dataset. Next, we apply non-negative matrix factorization (NMF) (Lee and Seung 1999) to identify separate behavioral segments and then to align these with customers segmented by demographics. The combination of the behavioral and demographic segments becomes the basis for the final integrated customer segments. NMF has been used in the prior work for customer segmentation (Shi et al. 2015b) but not for persona generation beyond our earlier work.

4.1 Problem formulation: general settings

Explaining the shape of the required dataset, our approach begins with one matrix encoding customers' interactions with content. We first build a matrix representing customers' interactions with the online products. We represent by **V** the g×c matrix of g customer segments (G_1, G_2, ..., G_g) and c pieces of content (C_1, C_2, ..., C_c). The individual elements of the matrix **V**, V_{ij}, are any value that represents the behavioral interaction by customer segment G_i for content C_j. In the case of YouTube Analytics, for example, V_{ij} is a view count of a particular video and C_j from customer segment G_g. In the case of Facebook Insights, for example, V_{ij} is the total minutes a particular video is watched and C_j, from a customer segment defined by gender, age, and country, such as [Female, 25–34, Australia]. Besides these two examples, there are other options to show the interaction between customers and content pieces, such as likes, ratings, and subscriptions. Such options can be used if they provide a breakdown of statistics across demographic groups. However, we note that such detailed statistics on demographics are not provided for views by both YouTube Analytics and Facebook Insights.

A customer segment (G_g) interacts with the set of digital products (C_1, C_2,..., C_c), so, a customer segment is defined as a set of the touch points with the digital content collection. With this matrix (**V**) as the basis, we can identify the distinct customer behavior patterns, which can be a vector of any set of customer touch points. Once we have the matrix **V**, we discover the number of significant latent patterns by decomposing it; that will become the basis of the personas, explaining the persona's preference toward content in the next step.

Regarding the generalizability of our method, we do not have hard constraints. This method is generalizable and applicable across (1) data of diverse granularity and (2) any content category. For example, a customer segment, G_i, can be an individual customer if the data is available at that granularity and if there is no privacy worry. This implies that the research method is generalizable to customer-level data, as well as to the aggregate-level data that we use here. We use only one matrix that represents attention to each content item. This type matrix can be easily accessed through many current social media analytic tools. YouTube Analytics and Facebook Insights provide statistics of attention (e.g., view counts) from a certain customer segment, defined by age, gender, or country for each video and post. Also, beyond social media analytic tools, our approach can be applied to any domain where the matrix **V** can be defined. As an example, if an online store provides statistics concerning which customer segments purchase which products, V_{ij} can be defined as the number of purchases from a particular customer segment i for a particular product j. Our method can find personas of that retail store without any modification of the core algorithm that we present here; therefore, this algorithmic method is generalizable.

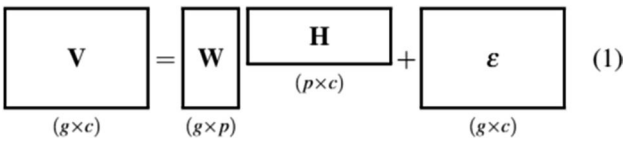

Fig. 1 Outline of matrix decomposition for identifying distinct behavioral segments and then impactful demographic segments

4.2 Non-negative matrix factorization to identify behavioral customer segments

Moving to our first research objective (recognize discrete customer segments based on behavioral interactions with online content posted on major online social media platforms), we use the social media data to identify discrete customer segments based on different behaviors. This process is quite challenging, as the customer statistics from most platforms are aggregated to preserve the privacy of the customers. Therefore, to isolate customer behavior patterns, the data must be disaggregated. For segmentation, we first experimented with k-means clustering (An et al. 2016a), but it was found ineffective because clustering, by definition, cannot break a given demographic segment into hidden behavioral segments. Thus, we turned to matrix decomposition techniques, specifically NMF, as outlined in (Jung et al. 2017), an approach used in other domains (Xu 2018). We conceptually present this matrix decomposition approach here.

Once we have the matrix **V**, as outlined above, the next step is to discover the underlying latent factors or the product behavioral patterns that become the basis of the customer segments. The matrix decomposition is presented graphically in Fig. 1.

As shown in Fig. 1, V is our $g \times c$ matrix of g customer segments (G_1, G_2, ..., G_g) and c contents (C_1, C_2, ..., C_c). When **V** is decomposed, **W** is a $g \times p$ matrix; **H** is a $p \times c$ matrix, and ε is an error term. In this case, p is the number of latent factors (behavioral patterns) that we can choose, which can control the resolution of the customer behavior patterns discovered. When we choose more latent factors, we get more fine-grained customer behavior patterns. The column in **W** is a basis for the segment, and the row in **H** is an encoding that consists of coefficients that combine with each basis and represent a linear combination of the bases. The resulting matrix decomposition equation is

$$\mathbf{V} = \mathbf{WH} + \boldsymbol{\varepsilon} \text{ or } V_{ij} = \sum_{k=1}^{p} W_{ik} H_{kj} + \varepsilon_{ij}. \quad (1)$$

In NMF, a column in **H** represents each of common content consumption patterns. The coefficient, H_{ij}, shows the importance of content, C_j, to explain the content consumption pattern, P_i. (i.e., distinct customer interaction pattern). As mentioned, **H** shows a set of distinct content consumption patterns represented by a linear combination of customer interactions with content.

4.3 Identification of representative demographics of behavioral customer segments

Moving to our second research objective (identify the discrete demographic customer segments associated with each of these behavioral customer segments), we find the most impactful customer demographic segments associated with the previously defined behavioral customer segments. For identifying the demographic customer segments, we take a two-step approach: (a) finding a set of representative demographic segments for each behavioral segment and (b) identifying the representative or most impactful demographics from this set. After decomposing the matrix **V**, we have the matrix **H** (containing the customer behaviors) and another matrix, **W** (containing the demographic groups).

First, we focus on **W** in Fig. 1. A row in **W** represents each customer segment consisting of different common behavior patterns. The coefficient, W_{ij}, is a relative proportion of a consumption pattern, P_j, in a customer segment, G_i (i.e., impactful customer demographic segment). A row in **W** represents how each customer segment can be characterized by different consumption patterns. A column in **W** shows how a distinct consumption pattern is associated with different customer segments. Thus, for each column, the customer segment with the largest coefficient can be interpreted as the most impactful customer segment for that corresponding pattern. A single behavioral segment is likely to have multiple associated customer demographic segments and vice versa. Thus, for each column, the customer group with the largest coefficient or weight can be interpreted as the most impactful customer demographic group for that corresponding pattern.

4.4 Integration of behavioral and demographics segments

For research objective three (integrate the associate behavioral customer segments and the demographic customer segments), we take a two-step approach of (a) finding a representative customer behavioral segment, as outlined above, and then (b) identifying the representative demographics of this group. Determining the demographics of the representative customer segments depends on how the customer segments are defined in **V**, the most efficient way is to use the data broken down by demographics when building **V**. For instance, if **V** has a row mapping into a group defined as [age group, gender, country], then it is trivial to find a segment's representative demographics. Social media analytics tools often provide customer statistics in a format that we can leverage for both demographic customer segments and for persona profile descriptive snippet (i.e., a short textual phrase describing a persona). Once we have identified the column in **W** with the largest coefficient, we then select that demographic grouping from our data set. For example, YouTube provides a demographic classification of 2 genders × 7 age groupings × 249 countries (3486 possible demographic groupings) per video, which is the ceiling of possible demographic segments that can be addressed.

5 Data collection

To develop and implement our approach, we leverage customer data from AJ+, an online news channel from Al Jazeera Media Network. In the highly competitive online news industry, understanding customers is notably important to both increase the consumption of digital content and to get relevant and noteworthy information to the readers that may be impacted by the news events. Two common goals for many organizations are to increase digital content consumption and to enhance the facilitation of digital content interaction with customers. Specifically, with the news industry, prior studies point out considerable differences between production and consumption patterns (Abbar et al. 2015), as the online news industry is competitive and fluid (Abbar et al. 2015; Kwak and An 2014). Therefore, in the news area, as with many other verticals, a proper understanding of customers is critically important, and this is an issue that online customer data can address (Mao and Zhang 2015; Shuradze and Wagner 2016).

We focus on the AJ+ YouTube channel[1] as the aggregated customer statistics data source, which we use as a proof of concept for our customer segmentation research, reserving future analysis of Twitter, Facebook, and other social media platforms for future work. However, the technique is generalizable to any online platform that provides aggregate customer statistics. The primary reason to focus on YouTube is that the analytics interface gives detailed statistics for every video. We do not lose generalization in showing proof of concept using a single media account because we use the data offered by YouTube, which has a universal format for all individual YouTube channels. In other words, our approach does not use any AJ+ dependent features; instead, we use the representative data that all YouTube accounts

[1] https://www.youtube.com/channel/UCV3Nm3T-XAgVhKH9jT0ViRg.

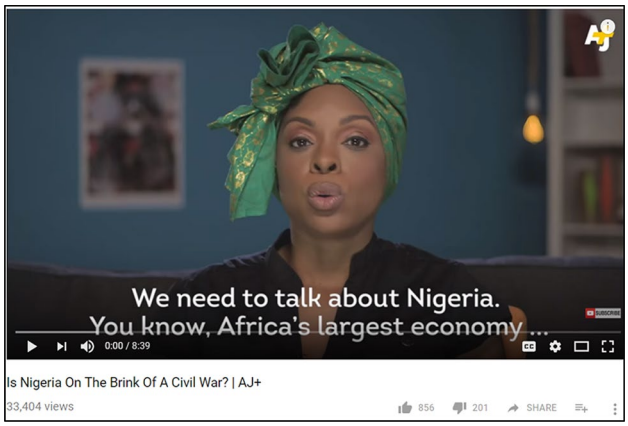

Fig. 2 Example of YouTube video from the AJ+ YouTube channel, with number of views

have. Also, the approach is transferable to other platforms, such as Facebook, that have identical or similar data variables (i.e., age group, gender, and location).

AJ+ is natively digital content platform, meaning that it was designed from the ground up to service news in the viewer's medium with no redirect to a website. AJ+ is based mainly on platforms. Therefore, digital content is specifically designed to be viewed on the Facebook Newsfeed, YouTube Channel, or Twitter Timeline depending on the readers who are most active on those platforms. As an example of an AJ+ YouTube video, see Fig. 2, noting specifically the number of views.

For the owner of the channel, the YouTube API provides analytics data for each video product and various customer profile data, (e.g., gender, age, country location, and which site the customer comes from), although at an aggregate level. Therefore, individual customer data is not provided. Via the YouTube API, we collect the detailed record of product views by country, gender, and age group for each of AJ+ video. In the research presented here, we focus on view counts due to their high volumes. A customer group is defined by gender, age, and country, such as [Male, 18–24, India]. We note that this detailed data breakdown is accessible with YouTube channel owner's (i.e., AJ+) permission. In summary, we collect data from 4,320 video products produced from June 13, 2014 to July 27, 2016. Collectively, these videos have more than 30 million views from customers in nearly 200 countries at the time of the study. Being quite robust, the YouTube analytics interface provides, for each video of the AJ+ channel, customer profile attributes (e.g., gender, age, country location, and which site the customer comes from) at an aggregate level. We use these customer and video attributes to explore if information dissemination can identify meaningful customer segment values based on video content interaction and on related demographics provided by YouTube. One can access the data in the YouTube analytics interface by the YouTube APIs[2]. The parameters we use for this research are listed below. There are various video KPI metrics; however, we only focus on *viewCount* (the number of views) in this research.

- Customer attributes

– *ageGroup* YouTube viewers are classified into seven age categories (13–17 years, 18–24 years, 25–34 years, 35–44 years, 45–54 years, 55–64 years, and 65 years and older).
– *gender* YouTube viewers are classified as either male or female, so there are two possible categories for a customer.
– *country* YouTube uses the two-letter ISO-3166-1 country code index to classify where viewers are from, with 249 current officially assigned country codes at the time of this study.

- Video attributes

– *viewCount* YouTube provides the number of views per video.

6 Results

Given our dataset, we define a customer segment as a unique combination of (*country, gender, ageGroup*). So, with two gender groups, seven age groups, and 249 countries, we have an upper limit of 3486 customer segments. (i.e., $2 \times 7 \times 249$). In actuality, our data set has 2214 customer segments, as the data has customers from 190 unique countries—we exclude as non-impactful those countries in which total view counts of 4320 videos are less than 1000 and those countries for which not all age groupings are represented.

6.1 Exploratory analysis of AJ+ YouTube data

We begin by presenting some of the overall statistics from the AJ+ YouTube channel data. Due to business concerns, we do not provide the exact absolute numbers, instead providing percentages only. The AJ+ customer population is worldwide with the top three countries, in terms of viewership, being Canada, Great Britain, and the United States (US), with each representing 2.44% of total viewership in terms of the number of unique videos watched. Regarding

[2] https://developers.google.com/youtube/analytics/.

the total number of views, the US is the largest customer market segment with about 49.4% of video views. Although AJ+ was designed to target the US market, it is interesting to note that most viewers come from outside the US, making it challenging to have a comprehensive understanding of the customer base.

Concerning customers' gender and age distribution, 20.9% of viewers were female with 79.1% being male. YouTube views are classified into multiple age categories (13–17 years, 18–24 years, 25–34 years, 35–44 years, 45–54 years, 55–64 years, and 65 years and older). As AJ+ is designed to target young generation by adopting platforms and that our data comes from the YouTube platform, it is logical that young adult males is the biggest segment.

For behaviors, some videos show a worldwide appeal, with 100 videos being viewed in 100 or more countries. Conversely, in the dataset, there were about 100 videos that were viewed by customers from five or fewer countries. In terms of the actual number of views, the viewership counts per individual videos follow a power law distribution with a small number of videos being viewed a lot and a large number of videos being viewed a small number of times. This finding is not surprising, as such skewed popularity of videos is one of the well-known characteristics of viewing behavior on YouTube (Cha et al. 2007).

6.2 Research objective one results—identification of customer behavior segments

To begin the decomposition, we first develop a matrix representing customers' interaction with the online content products. The matrix's columns are the online products, in this case, the AJ+ videos [e.g., c contents (C_1, C_2, ..., C_c)]. The matrix's rows are the customer segments or customer demographic segments (e.g., g customer segments [G_1, G_2, ..., G_g)]. Therefore, the matrix describing the association between customer segments and contents is denoted by **V** the g × c matrix of g customer segments or customer demographic segments and c contents. The element of the matrix **V**, V_{ij}, is any statistic that represents the one interaction or set of interactions of the customer group G_i for content C_j. In the research presented here, the customer interaction element is *viewCount*. Using this matrix approach as the basis, we can decompose (i.e., separate into simpler components) the overall matrix **V** into two matrices: **W** and **H**. The matrix **W** encodes an association between customer segments and behavioral customer segments (i.e., latent content interaction patterns), and the matrix **H** encodes an association between behavioral customer segments and pieces of content. The resolution in finding customer segments can be adjusted by the number of columns in **W** or that of rows in **H**. To sum up, once we have the matrix **H**, we discover the underlying latent patterns, which describe the customer interaction with content, and that will become the basis of the customer demographic segments in the next step.

Although one can present as many behavioral segments as the data contains, cognitive limits of the end users of the results pose a restriction; it is not purposeful to show them hundreds of customer segments. As the number of segments in our work is strongly tied to user experience and use, it is not best to compute the optimal number of segments in the matrix. Even if the optimal number is a large number, it is not good for persona creation because that number is too big to effectively employ in daily practice. For purposes of demonstrating the results in this manuscript, we present six customer behavioral segments in Table 1; although using NMF, we can generate as many segmentations as desired. In fact, this is the only parameter required, the number of segments, by NMF.

6.3 Research objective two results—identification of customer demographic segments

Moving to our second research objective, we identify the most impactful customer demographic segments associated with the previously defined behavioral customer segments. After decomposing the matrix **V**, we have the matrix **H** (containing the customer behaviors) and another matrix, **W** (containing the demographic groups). Each row in **W** represents how each customer demographic segment can be characterized by different consumption patterns. The columns in **W** show how a common consumption pattern is associated with different customer segments. A single behavioral segment can, possibly, have multiple associated customer demographic segments. Thus, for each column, the customer group with the largest coefficient or weight can be interpreted as the most impactful customer demographic group for that corresponding pattern. Although one can present as many behavioral segments as the data contains, cognitive limits of the customers of the system pose a limit; it is not purposeful to show them hundreds of customer segments. In terms of populating **W**, the YouTube Analytics interface provides demographic percentages (based on gender and age) of viewing for each video by country (as shown in Table 2). Using this data, we can populate the demographic attributes of **W**, which we then associate with **H**, which contain the customer segments.

6.4 Research objective three results—integrating customer behavioral and demographic segments

Moving to our third research objective, we integrate the most impactful customer demographic segments associated with the previously defined behavioral customer segments. A single behavioral segment is likely to have multiple associated

Table 1 Results of NMF for matrix **H** showing six customer behavioral segments and associated weights for twenty of the videos

Videos	Customer behavioral segments (1 through 6)					
	1	2	3	4	5	6
ElyepfzS1bU	0.73822	0.04768	0.210637	0.204819	0	7.189821
DhMwc2FyGC0	0.019681	0	0.013142	0.148262	0	1.306572
C32VsZRaAzg	0	0.001238	0.048973	0.064777	0	0.415397
-s2CJm8hFc0	0.343069	0.020808	0.105554	0	1.106555	3.619799
z-Y31PEeIyo	0	0	0.085692	1.410283	0	0.998431
xvLm2kdbhkc	0	0.011591	0.044462	0.064921	0	0.634569
xt6UwfmoLb4	0	0.001204	0	0.028282	0	0.254913
xsI17XsJdCs	0.447489	0	0.043879	0.175264	1.536789	2.461493
vF0FskqoO10	0	0.004158	0.061832	0.058873	0	0.937093
v-g3LYBX5ws	0	0.003459	0.018849	0.05622	0	0.284175
ubrZpTBAXgg	0.145552	0	0.04386	0.107746	0	2.019695
uYwTjFBbasw	0	0.002438	0	0.044098	0	0.438586
s753-qB_fgs	0	0.022296	0.243359	1.164752	0	1.061365
rUWnQ8cEqT8	0	0	0	0.087995	0.058039	0.190359
rTZPJYqbQ6M	0	0.003044	0.026096	0.16786	0	0.525854
qVzTmbE6G60	0	0.0158	0.077921	0.36179	0	1.142301
oUs0YBLJgHc	0	0.000693	0.017842	0.07145	0	0.238511
nuGGbsLq7kA	0.062496	0	0.042316	0.22834	0	1.269035

The entire matrix not shown due to space limitations

Table 2 Results videos and viewing by demographics used for the demographic data in matrix **W** showing two 14 age and gender customer demographic segments per video by country and associated view counts for 20 of the videos

Video	Country	Video view	Age and gender category 1	Age and gender category 2
0IBZ1PxmOAw	US	234,470	Age 13–17, female, 2.2	Age18–24, female, 9.9
U3ey4xDoKtU	TW	227,532	Age13–17, female, 1.2	Age18–24, female, 7.6
qRYW_I60xoM	US	209,472	Age13–17, female, 2.2	Age18–24, female, 10.4
Hyv39DxQlFM	US	169,955	Age13–17, female, 1.1	Age18–24, female, 7.4
5MQjsXRqLmA	RU	167,995	Age13–17, female, 2.2	Age18–24, female, 10.4
U3ey4xDoKtU	MX	144,615	Age13–17, female, 1.1	Age18–24, female, 4.9
e_EeVleqo58	PK	139,882	Age13–17, female, 0.5	Age18–24, female, 2.6
0IBZ1PxmOAw	FR	105,568	Age13–17, female, 1.8	Age18–24, female, 11.7
izdfnHBMwSs	CA	99,361	Age13–17, female, 1.2	Age18–24, female, 7.5
tk2hOYUpKVI	ID	96,895	Age13–17, female, 3.0	Age18–24, female, 14.7
jL-li8ZJGFo	US	96,563	Age13–17, female, 1.2	Age18–24, female, 6.2
v8RGUDHEIJo	CA	93,729	Age13–17, female, 1.6	Age18–24, female, 9.1
qHtJSyWK0zw	IN	91,794	Age13–17, female, 0.5	Age18–24, female, 4.9
z-Ize9i9Zd0	PK	91,624	Age13–17, female, 1.2	Age18–24, female, 3.7
dOexgirwN0w	US	89,895	Age13–17, female, 6.0	Age18–24, female, 13.1
XrrNzH_7kjY	US	89,203	Age13–17, female, 2.2	Age18–24, female, 7.5
e_EeVleqo58	SA	88,349	Age13–17, female, 0.3	Age18–24, female, 0.9
aiVxyLb1hJA	CA	87,135	Age13–17, female, 0.7	Age18–24, female, 7.5
h2TPlxBIvOQ	CA	86,711	Age13–17, female, 1.6	Age18–24, female, 5.8
5UjIqwf9fyk	NL	82,949	Age13–17, female, 0.4	Age18–24, female, 7.1

The entire matrix of both videos and demographic segments not shown due to space limitations

demographic customer segments. Thus, for each column, the customer group with the largest coefficient or weight can be interpreted as the most impactful customer demographic group for that corresponding pattern. Although one can present as many behavioral segments as the data contains, cognitive limits of the system's customers pose a restriction; it is not purposeful to show end users of analytics information of hundreds of customer segments. Therefore,

Table 3 Six customer behavioral segments presented with the top six associated customer demographic segments

Customer behavioral segment	Possible customer demographic segments			
No.	Country	Age	Gender	Weight
1	**United States (US)**	**18**	**Male**	**1342.20**
	US	25	Male	816.23
	US	13	Male	425.73
	US	35	Male	416.56
	US	18	Female	303.16
	US	45	Male	267.99
2	**US**	**25**	**Male**	**3139.72**
	US	18	Male	1,199.25
	US	35	Male	993.53
	CA	25	Male	332.51
	US	45	Male	305.59
	US	25	Female	294.09
3	**Pakistan (PK)**	**18**	**Male**	**244.00**
	PK	25	Male	127.04
	PK	13	Male	124.60
	PK	35	Male	39.70
	PK	18	Female	22.39
	PK	45	Male	16.21
4	India (IN)	18	Male	333.47
	IN	25	Male	332.65
	SA	25	Male	245.20
	PK	25	Male	220.57
	PK	18	Male	167.59
	IN	35	Male	114.54
5	**US**	**25**	**Female**	**445.14**
	US	18	Female	403.55
	US	35	Female	245.49
	US	45	Female	186.23
	US	35	Male	183.67
	US	45	Male	166.03
6	**IN**	**25**	**Male**	**162.59**
	IN	18	Male	145.64
	GB	25	Male	15.75
	CA	25	Male	11.80
	GB	18	Male	9.37
	IN	18	Female	8.76

The demographic segment with the most weights is bolded, which we use as the representative demographic segment

we constrain the number of segments shown to end users. Table 3 shows six demographical customer segments for each of the six behavioral customer segments.

In Table 3, we present five customer demographic segments identified via our matrix decomposition approach to discover distinct customer behavioral segments based on online content interaction (see Column 1). Then, for each behavioral pattern, Table 3 displays the top customer demographic segments associated with each of these five behavioral segments (see Columns 2 through 5). Our decomposition approach calculates a weight for each of these demographic segments, assigning a higher weight to the most impactful (i.e., largest) demographic customer segments.

The results in Table 3 also demonstrates how our approach using NMF is more effective than using clustering methods such as K-means++ in finding meaningful customer behavioral segments with representative demographic groups. In our previous work (An et al. 2016a), we applied the clustering method (K-means++) to YouTube data to find a set of groups that share the common video consumption patterns. When using K-means++, all 14 (7 age groups × 2 gender groups) US demographic groups were clustered as one group. The results by K-means++ were technically correct. However, such clustering results are not practical in actual use because they miss hidden behavioral segments within each of demographic group. The US is the biggest customer segment in our dataset, and thus even within one demographic group (e.g., US-25-Male), there exist a few different video consumption patterns. Since our data is aggregated, clustering methods such as K-means++ cannot capture such behavioral differences. Unlike clustering methods, decomposition methods such as NMF can "decompose" the aggregated data, identifying subtle behavioral differences among the demographic groups. As a result, NMF method results in presenting three US demographic groups as representative behavioral segments while K-means++ results in having one US group. Considering that US customers are the majority in our data, having three groups is more reasonable than having one group.

In Table 3, there are at least two findings that are apparent upon analysis. First, there is a predominance of male segments. Second, there is a clustering effect by gender, and then age, and sometimes location. For example, in the first behavioral segment, we see that the top three demographic segments are all young males from the United States. This opens up an interesting research question of how granular one needs to get for each of these demographic segments, as there is apparently little behavioral difference when segregating this portion of the population into three different age brackets.

6.5 Research objective four results—integrating customer behavioral and demographic segments

We believe that there are many possible use cases for employing behavioral and demographic segments, both separately and in conjunction by linking the behavioral segments to demographic segments. Here, we present one possible use case, using the research results to automatically

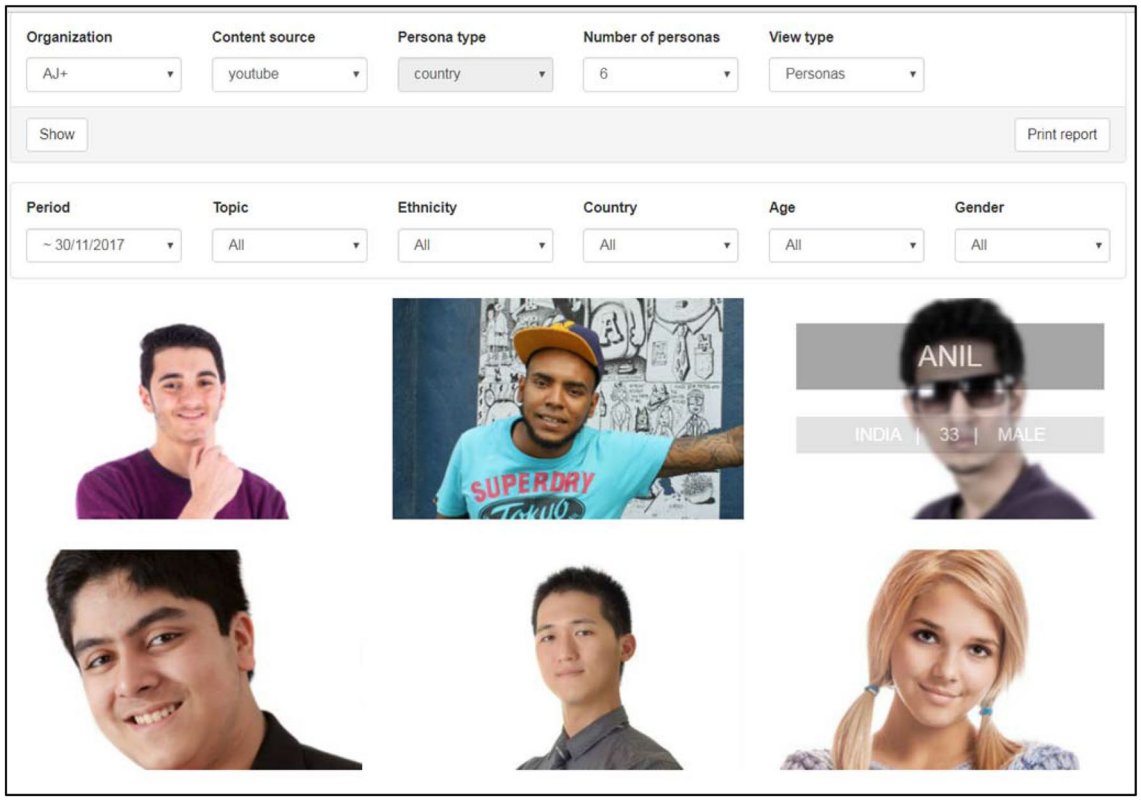

Fig. 3 Screenshot of the Automated Persona Generation System generating six personas based on YouTube data. Note the images for each persona and the demographic information that appears on the cursor rollover of one of the images

generate personas using social analytics data to first isolate customer segments, both behavioral and demographic. We have developed a system that automatically collects aggregated data and that decomposes it using the method outlined above. We then turn the customer segments into rich personas by adding personality attributes to each, as outlined in other work (An et al. 2017; Jung et al. 2017). The result of this is that we automatically generate personas based on actual customer data from online platforms, a significant evolution of persona creation research, which can be used standalone or in conjunction with other persona creation methods.

As shown in Fig. 3, with an example of six customer segments as the bases of the personas (fictive people based on real data), the system presents a demographically appropriate image, name, country, age, and gender for each persona. This demographic data is first derived from the social media data, i.e., gender, country, and age. Using this information, the system then accesses backend databases selecting gender, age, and country appropriate images and names.

The demographic information is displayed when the cursor hovers above a persona image. When one of the persona images is clicked, the corresponding persona description is displayed, as shown in Fig. 4.

In choosing the number of personas to generate (or display via the system), we give end users the flexibility to choose the number of personas generated. The number should be much smaller than the number of total groups ($|\mathbf{G}|$) and that of contents ($|\mathbf{C}|$) because the condition for NMF is $|p| << \min(|G|, |C|)$. In the case of AJ+, customers can choose any number of personas from 5 to the order of 15. However, the cognitive load of hundreds of personas may make personas unusable as it is difficult to make sense of so many sub-groups of the audience; therefore, in practice, a smaller number is more reasonable, although in theory, the method and resulting system can generate as many personas as desired and as indicated by the data.

6.6 Scalability of discovering behavioral segments in empirical settings

Prior to moving on to the quantitative evaluation of customer segmentation methodology, we show the scalability of discovering behavioral segments in various empirical settings. Our approach to building personas can be divided into two parts: NMF and refinement of behavioral segments by adding personality, such as name, photo, etc. The latter is a simple task of searching the database and thus can be

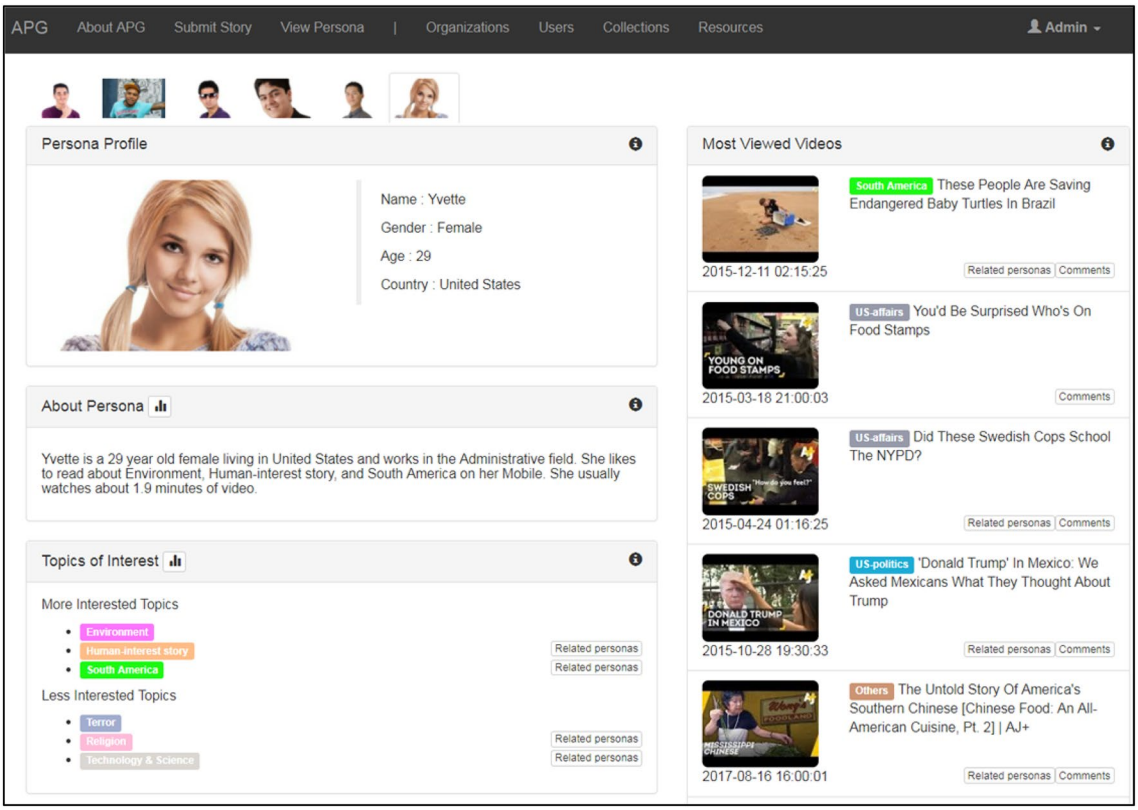

Fig. 4 Screenshot of a persona description that is automatically generated from social media analytics and contains both behavioral and demographic customer segmentation data

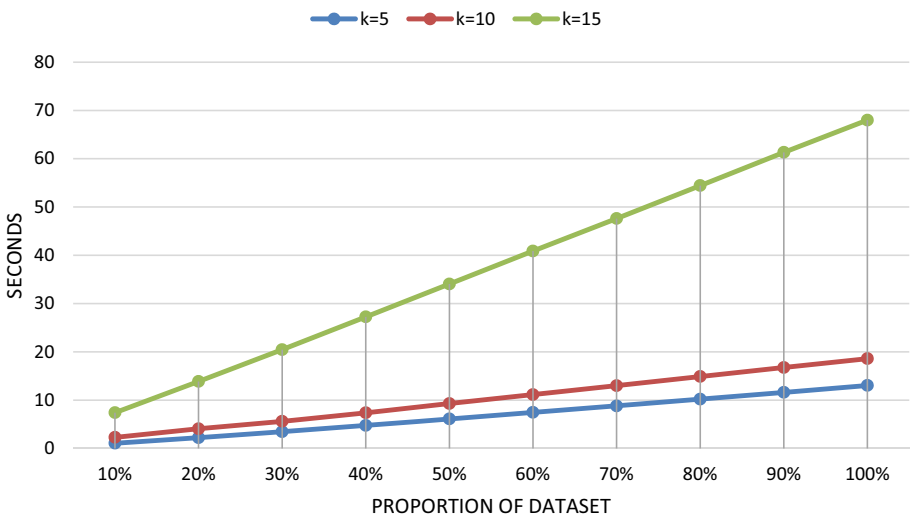

Fig. 5 Elapsed time (seconds) of running NMF with different sizes of the data and different numbers of the behavioral segments

done in $O(k)$ where k is the number of segments. The former can have various time complexities based on which implementations can be used. Here we use Python scikit library, which has $O(gck)$ where g is the number of groups, c is the number of contents, and k is the number of segments. Thus, we measure the elapsed time of only the NMF part varying the size of the input matrix. Considering the distribution of the real traces, we sample the original AJ+ matrix with 10% intervals and measure the elapsed time of NMF with a commodity setup (a laptop with Intel i7-3770 CPU and 12.0 GB memory).

Figure 5 shows the elapsed time of running NMF with different sizes of the data and different numbers of the behavioral segments. We run the approach 500 times for

each configuration and compute the average elapsed time for each. As shown in Fig. 5, the elapsed time for running NMF linearly grows with the size of the data and remains under minutes with a commodity laptop. Optimization and distributed computation might shorten the elapsed time more. Considering that our matrix is built from one entire news media outlet, our approach can be applied to other comparable social media accounts.

6.7 Consistency in the resulting matrices of NMF

As the NMF is an approximation algorithm, the resulting matrices might be changed by parameters, such as initial values or numerical solvers. Also, algorithms might not converge in a specific situation (Lin 2007). To show the resulting matrices' consistency in an empirical setting, we tested our algorithms with different settings and compared the results.

We ran our algorithms with different initializations, which are (1) random, (2) Non-negative Double Singular Value Decomposition (NNDSVD), (3) NNDSVD with zeros filled with the average of the original matrix, and (4) NNDSVD with zeros filled with small random values, provided by Python scikit learn library. Also, for each initialization, we test different numerical solvers, which are Coordinate Descent solver and Multiplicative Update solver. As a result, we have $4 \times 2 = 8$ different configuration of initializations and numerical solvers.

For each configuration, we ran it 500 times for each experiment and extracted representative groups for k (the number of personas) equaling 5 and 10. We report the results from two perspectives. One is the consistency within a given configuration. Running 500 times of NMF, we obtain the same representative groups for all of them. Not a single run with different representative groups exists. The other is the consistency across the configurations. When $k = 5$, we find three representative groups appear 4000 times (500 times $\times$ 8 configurations). The remaining two representative groups for each configuration are different across the configurations. However, their variations are quite limited; four different groups appear 3500, 1500, 2000, and 1000 times. When $k = 10$, we find five representative groups appear 4000 times, and for remaining spots, seven different groups appear 3500, 3500, 3500, 3000, 2000, 500, and 500 times. Of course, there are methodologies to pick the best configuration among them by minimizing the difference that is measured by Frobenius norm, between the product of $\mathbf{W}$ and $\mathbf{H}$ and the original matrix $\mathbf{V}$. Nevertheless, most of the results are quite stable and converge with the empirical data.

7 Quantitative evaluation of customer segmentation methodology

To quantitatively evaluate our approach to customer segmentation, we conduct two analyses using a ten-fold methodology on the data set. The two analyses are (a) predicting the most impactful customer demographic for a given customer behavioral segment and (b) predicting new video views by demographic.

7.1 Predicting customer segment interest in new content and number of video views

One of the benefits of using NMF for generating customer segments and personas is a clear association, represented in $\mathbf{H}$ ($p \times c$), between the customer segments' interest and non-interest in specific digital content. Beginning with this association, we can identify content, H_n, that a given customer segment might be interested in even before content publication.

For the problem of predicting interest in new content, the most intuitive solution is to find similar content that has already been published relative to the new content and assume that the level of interest in similar content will remain the same by a given customer segment. To compute the similarity of content in a robust way, we define content features. The features can be anything: topics, length, mood, color, price, and so on. Formally, we define a matrix, F ($c_{contents} \times f_{features}$), capturing the features of the content. We then can derive another matrix, $K(p_{personas} \times f_{features})$, that represents an association between a customer segment and content features:

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

where k is a kernel function. Thus, we can rewrite Eq. (2) with some appropriate mapping function φ:

$$K = \varphi(\mathbf{H})\varphi(\mathbf{F}). \tag{3}$$

For computational simplicity, we assume $\varphi = \mathrm{I}$. In other words, the interest in content is the sum of the interest in its features. Then, we can get a direct multiplication of two matrices:

$$K = \mathbf{HF}. \tag{4}$$

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

$$\mathbf{H} = \mathbf{K}\mathbf{F}_{\text{right}}^{-1}, \tag{5}$$

where $\mathbf{F}\mathbf{F}_{\text{right}}^{-1} = \mathbf{I}$.

The representation of $\mathbf{H}$ in Eq. (5) guides us on how to predict $\mathbf{H}_n$. For new content, we can define $\mathbf{F}_n$ that

represents new content and their features. By substituting $\mathbf{F}_n$ into Eq. (5), we can get $\mathbf{H}_n$:

$$\mathbf{H}_n = \mathbf{K}(\mathbf{F}_n)^{-1}_{\text{right}}. \tag{6}$$

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

$$(\mathbf{F}_n)^{-1}_{\text{right}} = \boldsymbol{F}_n^T(\boldsymbol{F}_n\boldsymbol{F}_n^T)^{-1}. \tag{7}$$

Equation (7) is valid when $\mathbf{F}_n$ has linearly independent rows ($\boldsymbol{F}_n\boldsymbol{F}_n^T$ is invertible). If not, we split a set of new content products into several sets so that $\mathbf{F}_n$ of each set has linearly independent rows. This procedure avoids losing the method's generality.

By combining Eqs. (6) and (7), we write Eq. (8), representing the association between customer segments and new content:

$$\boldsymbol{H}_n = \boldsymbol{K}\boldsymbol{F}_n^T(\boldsymbol{F}_n\boldsymbol{F}_n^T)^{-1}. \tag{8}$$

The key of Eq. (8) is that $\mathbf{K}$, the matrix representing an association between customer segments and features, does not need to be changed for newer content because $\mathbf{K}$ depends on content features, not the content itself.

This is an application and an advantage of our customer segmenting methodology relative to other limited approaches that have been attempted for online data-driven customer profiling methods. In addition to providing segments of their customers, our approach also identifies the target customer segment for new content once its features are selected and measured. The content creators then have an opportunity to refine their content prior to its publication to more directly appeal to the customers they want to target.

By combining Eqs. (1) and (8), for new content, we get $\mathbf{V}_n$:

$$\boldsymbol{V}_n \cong \boldsymbol{W}\boldsymbol{H}_n = \boldsymbol{W}\boldsymbol{K}\boldsymbol{F}_n^T\left(\boldsymbol{F}_n\boldsymbol{F}_n^T\right)^{-1}. \tag{9}$$

Similar to Eq. (8), it is possible to predict the views of new content by customer segment based on a content feature of the new product.

7.2 Experimental setup for evaluation

Using this approach, we first define a training and a testing data set. We divide all 4323 videos, ordered by publishing date, into 10 slices. Among the 10 slices, we use the 10th slice as our testing set, which is the latest 432 videos. Then, for training, we use some of the remaining slices to consider the recency and their expressive power given that these videos represent the most current audience preferences. Although there is a general belief that more training data leads to better prediction performance in machine learning (Brownlee 2016), in our case, more videos to train the model is not necessarily helpful because the customer base might change and evolve over time. In such cases, older data might not reflect the behavioral patterns of the current customers.

To better understand how the size or the recency of the training set affects the prediction performance, we iteratively run an experiment with a varying number of slices in the training data from one (the most recent) to ninth (the oldest). For clarity's sake, we use the percentage of the testing data to the whole instead of the number of slices; $N = 10\%$ means the ninth slice only, and $N = 30\%$ means the 7th, 8th, and 9th slices. We get nine different sizes of the training data sets by changing N from 10 to 90%, with an offset of 10%. For each training set, we construct a matrix $\mathbf{V}$, and by applying NMF, we get a matrix $\mathbf{W}$ and $\mathbf{H}$. Then, we build a Latent Dirichlet Allocation (LDA) (Blei et al. 2003) topic model for each training set to construct a matrix $\mathbf{F}$ and $\mathbf{F}_n$. Once we construct these five matrices, we estimate the view counts of new videos for demographic groups, $\mathbf{V}_n$ (g groups $\times$ n videos), according to Eq. (9).

7.3 Measure of evaluation—Kendall's coefficients

For each of the new videos in the 10th set, we rank the demographic groups based on weight values in $\mathbf{V}_n$. We compare this ranking with the true ranking of the groups computed from the real view counts of that video by Kendall rank correlation coefficient. Since we have 432 test videos, we have 432 Kendall's coefficients (t) for each experiment run. For evaluation, we first use the mean of those 432 Kendall's coefficients; a higher coefficient means a method performs better in ranking demographic groups for a video, which represents how well a method perform in finding. Second, we present how many of the 432 test cases have statistically significant results. The higher the number of significant cases, the better the method performs in identifying demographic groups with a higher view of a given video.

For comparison to a baseline, we employ two other models: (1) a random model and (2) a collaborative filtering (CF) model. The random model ranks groups randomly for a new video. The CF model computes the average view counts of each demographic group and uses them for any new video, as CF-based recommending system assigns an average behavior of customers for the new content that would represent what one would consider a standard web analytics approach (Bowden 2009).

7.4 Results of evaluation

Figure 6 shows the result of the experiment: (a) the average t of cases where the result is statistically significant ($p < 0.05$) and (b) the number of those significant cases in each experiment. The inferior performance of the random

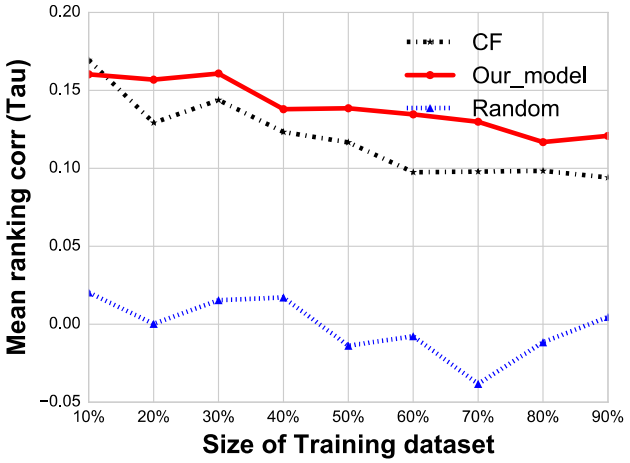

Fig. 6 The result of predicting the ranking of demographic groups by Random, CF, and our model. *Y* axis value is the average of Kendall's rank correlation coefficient of our 432 test cases. *X* axis value is the percentage of the dataset used for training. Ten percent means we use the last 10% of the dataset before the test dataset

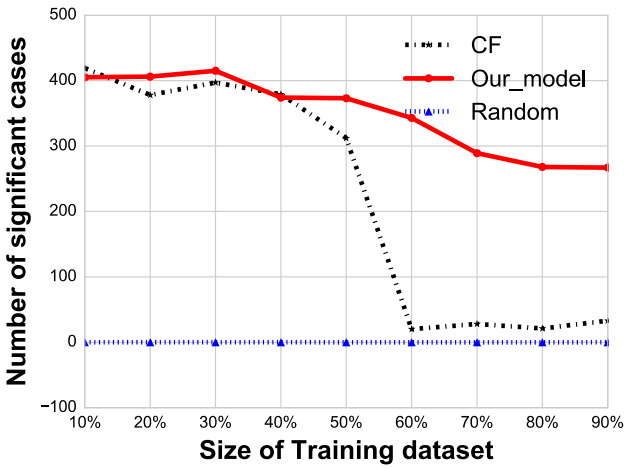

Fig. 7 The number of significant cases when testing Kendall's Rank Correlation coefficients among 432 test cases by Random, CF, our model. *Y* axis value is the number of cases where Kendall's rank correlation coefficients are significant among 432 test cases

model proves that the view counts from each segment for the videos are far from the random construction. In fact, there are not many significant cases for the random model (see Fig. 6), and the average *t* for the random model is near 0.0 for any size of the training set.

Figure 7 shows that our model outperforms the CF-based model in ranking the groups for new videos when *N* is 20–90%, and it shows comparable performance when *N* is 10%. These results demonstrate that our customer segment prediction performs very well at forecasting interest in new content by customer segment and that our approach outperforms, from 20 to 90%, the widely used CF approach, even with a set of limited features.

Also, for the CF-based model, the number of significant cases strikingly drops, from 412 significant cases to 18, when *N* is greater than 40% (Fig. 6). From our stability analysis, we show that the channel encountered a sudden change in their consumer customer in that period. The CF-based model is not robust for such sudden changes, resulting in having no significant cases when $N > 40\%$. The average *t* of the CF model would show a significant drop when $N > 40\%$ if we plot the average *t* of all cases. In contrast, our customer segmenting approach is robust enough to adapt to the changes of the audience, as shown by the number of significant cases in Fig. 6, even when *N* varies from 10 to 90%.

8 Evaluation of distinctiveness of personas by varying the number of personas

In this section, we offer insight into the relationship between the number of personas and their distinctiveness. Two contradictory patterns can possibly emerge. On one hand, the personas can become more distinctive when the number of personas increases. As a higher number of personas are discovered, subtle differences are likely to be captured by different personas that otherwise would be subsumed into the same persona. On the other hand, the personas can overlap when the number of personas increases. This might happen when the number of actual customer segments is smaller than the number of personas. For an analogy, consider a set of red balls and blue balls. Then, a mixed segment of red and blue balls cannot be avoided if finding more than two segments.

We define distinctiveness of personas as the average distance between rows in **H**. More specifically, we consider each row in **H** as a c-dimensional vector and compute the average cosine distance between all the pairs of two vectors. The number of possible pairs here is ${}_pC_2 = p(p-1)/2$.

Figure 8 shows how the distinctiveness of personas with varying the number of the personas. Interestingly, the emerging pattern is neither the two possible patterns we mentioned above. Rather, it is a complicated combination of them. The distinctiveness first sharply decreases until four or five personas are found, and then it steadily increases and becomes almost stable.

This tendency highlights the difficulty in choosing the optimal number of personas in a real scenario. As online content is consumed by millions of customers and their grouped behavior is not as simple as a theoretical example, the number of latent patterns and their distinctiveness shows the complex nature rather than a simple positive or negative correlation. However, while a higher number of personas than four or five can capture fine-grained differences in

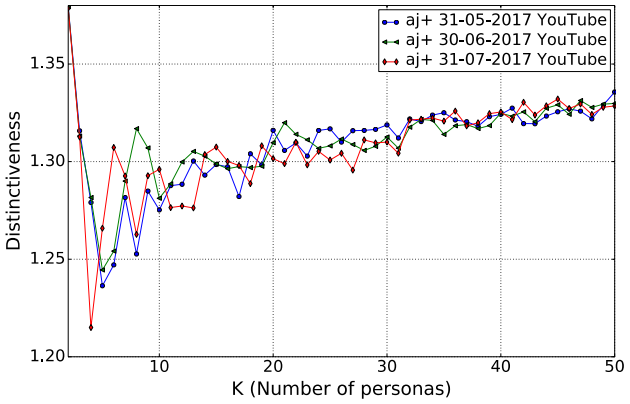

Fig. 8 Distinctiveness of personas with varying the number of personas for YouTube dataset

customer behavior, it increases the cognitive load to handle this information at the same time. Thus, in the real scenario, the number of personas should be carefully chosen by those who actually use the personas or customer segments in their work routine.

9 Discussion and implications

The results of our research demonstrate that social media data from the major online platforms is quite robust in identifying customer segments for both behavioral and demographic segments. We demonstrate several important results and implications concerning customer segmentation. Notably, the approach shows that one can use actual, real-time, aggregated online customer data at scale to identify meaningful customer segments and then can automatically generate personas from these segments. As such, this research addresses a previously open question of investigation and advances both segmentation and persona research by presenting an approach to use data at scale and to continually leverage this data to keep personas updated. Using our method, we do not need to vigilantly survey customers, who may or may not be actual consumers of our content. We can leverage online social media data from actual consumers to extract customer segments based on genuine customer data. Additionally, our method can be used to supplement qualitative methods of personas creations.

The major strength of our approach is that it benefits from actual customer data, reducing time and cost for generating both behavioral and demographic customer segments and providing a mechanism for linking these two types of customer groupings into coherently integrated segments. Also, in contrast to prior persona research, our research focuses on using data from major online platforms that are, in most cases, aggregated to some level. Prior work in using online data as the basis of personas used individual customer 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 customer segments and then to automatically generate rich personas. In this research, we also generate a relatively large number of personas. While prior work has recommended a small number of personas, this is not feasible or realistic for content, systems, or platform channels with millions of followers. This is a first step in focusing persona research on the needs of the designers and producers of online content. Our research focuses on the increasingly common situation of digital content creators that are distributing their content to an extremely large, heterogeneous customer base via major online platforms, which is, or is becoming, the de facto technologies of distribution. In this situation, the role of the personas is to identify customer tastes and interests, rather than in more traditional system interactions. However, we believe the approach could be applied in these situations also.

Although the research presented here leverages YouTube data, the method is transferable to most platforms, as the data collected is in similar aggregated format. Like YouTube Analytics data, Facebook Insights provide content consumption statistics from a certain customer segment, defined by age, gender, and country for each video and post. Unlike YouTube, Facebook Insight provides the total view time (total minutes of the video watched) of a video instead of the number of total views. We use two different API calls, "view_time_by_age_bucket_and_gender" and "view_time_by_region_id" to calculate the view time of each video for each customer segment. Once we have this data, then we create matrix V, and the following step is same as we did for the YouTube data. We note that there is no extra cost except collecting data for our system to process datasets from two different online platforms. In fact, we have implemented the approach using Facebook data for the same organizations with the resulting personas displayed in Fig. 9.

Yet, there are numerous research and development fronts that we are pursuing in the future to enhance the impact of this research. Given the reliance on streams of social media data, we could certainly implement direct access to the foundational customer data for the content creators, as suggested by (Faily and Flechais 2011) and also to persona campaigns, as suggested by (Judge et al. 2012), where updates concerning the personas are continually sent to the content creators. This feature may be important, as it appears that designers like continued access to the actual customer data, aside from the persona itself, to aid them in their design decisions (Judge et al. 2012). In this work, we use NMF as the basis for our persona generation. To identify customer segments, we are also investigating other prominent and advanced approaches besides NMF such as convergent NMF (Mirzal 2014), PCSNMF decomposition (Shi et al. 2015a),

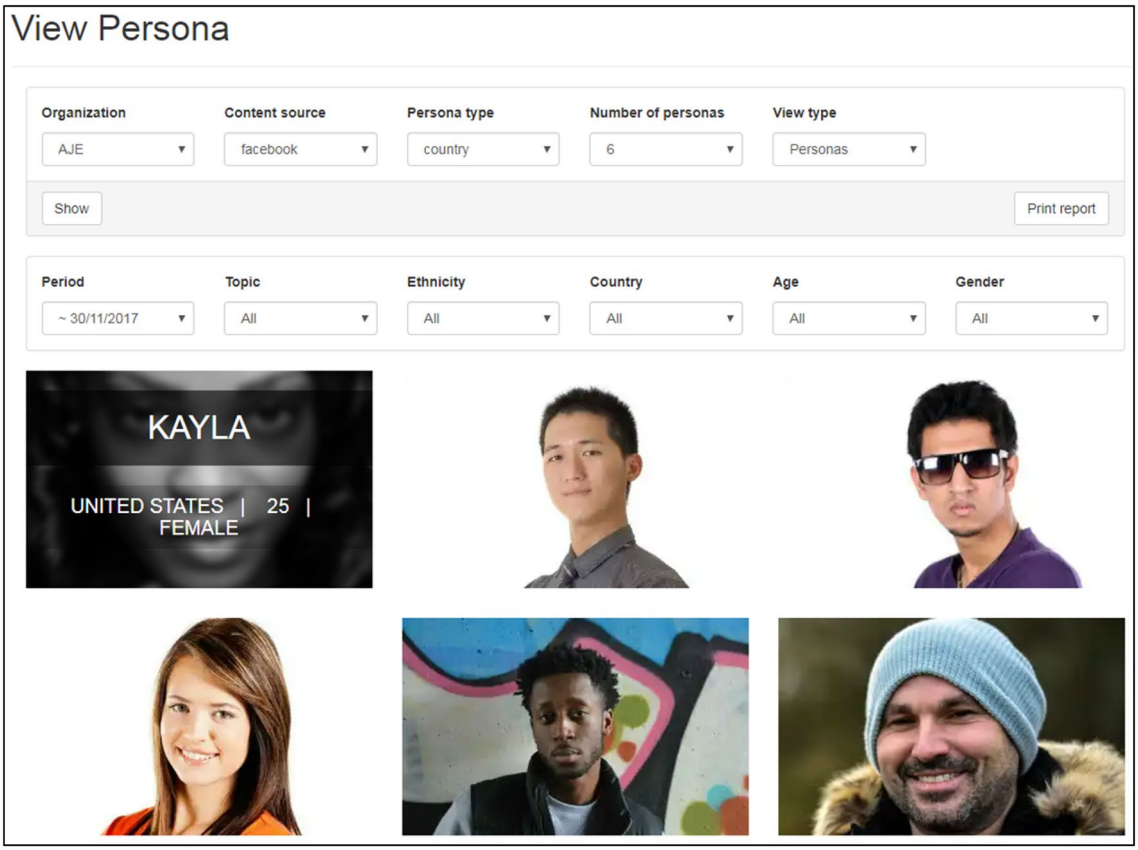

Fig. 9 Screenshot of the Automated Persona Generation System generating six personas based on Facebook data. Note the images for each persona and the demographic information that appears on the cursor rollover of one of the images

and rank-adaptive NMF (Shan et al. 2018). Also, we used the LDA topics for characterizing videos based on video headlines. However, there are many other candidate features (Zarrinkalam et al. 2018). Carefully selecting an increased number of features could improve the performance of our model. Most importantly, as mentioned above, we will conduct an in-depth field evaluation, such as that reported in (Dittmar and Hensch 2015), of the system with actual journalists, both producers and editors.

We consider this research is a starting point for leveraging behavioral and demographic customer segments from social media analytics data for a vast number of other applications and services with minimal manual efforts. If we can leverage additional rich information concerning a consumer, such as an ethnicity, socio-economic status, and precise location, our approach and results would become even more useful. For future research, we are exploring these avenues. For example, it might be possible to extract demographic information using shared links on Facebook (An et al. 2016a), via Twitter, or via Google + profiles. Links shared on Facebook could reveal information about the customers, such as socio-economic status, as the links reveal particular interests. There has been prior research showing that affluent customers visit more high-end luxury product websites, while budget-conscious customers visit price aggregation or discount websites. Thus, the socio-economic status of the consumer can be distinguished by the websites they visited (Odlyzko 2003). Other features, such as psychographics, political orientation, and brand affiliations could also be associated with the personas based on interest mapping.

10 Conclusion

In this research, we show that personas can be rapidly and automatically created from large-scale, aggregated customer data from major online platforms, resulting in personas that are based on behavioral data that reflect real people and created from sizeable data quantities permitting quantitate analysis. We evaluated our persona generation methodology, showing that our method generates actual and stable personas that are predictable. Although specifically focusing on digital content creators, our approach is flexible and resilient for application in a wide range of contexts where customer-centric data needs to be transformed into easy-to-understand representations for decision-making and customer insights.

Acknowledgements We thank the many journalists at Al Jazeera News Media Network for their collaboration in this research.

References

Abbar S, An J, Kwak H, Messaoui Y, Borge-Holthoefer J (2015) Consumers and suppliers: attention asymmetries. A case study of Aljazeera's news coverage and comments. In: Paper presented at the Computation + Journalism Symposium 2015, New York, NY, 2–3 Oct

An J, Cho H, Kwak H, Hassen MZ, Jansen BJ (2016a) Towards automatic persona generation using social media. In: 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), 22–24 Aug 2016, pp 206–211. https://doi.org/10.1109/W-FiCloud.2016.51

An J, Kwak H, Jansen BJ (2016b) 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, 29 Nov–2 Dec 2016

An J, Kwak H, Jansen BJ (2017) Personas for content creators via decomposed aggregate audience statistics. In: Advances in Social Network Analysis and Mining (ASONAM 2017), Sydney, Australia, 31 Jul–3 Aug 2017, pp 632–635

Antoniou A (2017) Social network profiling for cultural heritage: combining data from direct and indirect approaches. Soc Netw Anal Min 7:39

Beane TP, Ennis DM (1987) Market segmentation: a review. Eur J Mark 21:20–42

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Bonoma TV, Shapiro BP (1984) Evaluating market segmentation approaches. Ind Mark Manag 13:257–268

Bowden JLH (2009) The process of customer engagement: a conceptual framework. J Mark Theory Pract 17:63–74

Brownlee J (2016) Machine learning performance improvement cheat sheet. https://machinelearningmastery.com/machine-learning-performance-improvement-cheat-sheet/. Accessed 9 Apr 2018

Cha M, Kwak H, Rodriguez P, Ahn Y-Y, Moon S (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the 7th ACM SIGCOMM conference on Internet Measurement, pp 1–14

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

Chéron EJ, Kleinschmidt EJ (1985) A review of industrial market segmentation research and a proposal for an integrated segmentation framework. Int J Res Mark 2:101–115

Chiang M-F, Lim E-P, Low J-W (2015) On mining lifestyles from user trip data. In: Paper presented at the Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, Paris, France

Clarke MF (2015) The work of mad men that makes the methods of math men work: practically occasioned segment design. In: Paper presented at the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea

Cooil B, Aksoy L, Keiningham TL (2008) Approaches to customer segmentation. J Relat Mark 6:9–39

Cooper A (2004) The inmates are running the asylum: why high tech products drive us crazy and how to restore the sanity (2nd Edition). Pearson Higher Education, New York

Dharwada P, Greenstein JS, Gramopadhye AK, Davis SJA (2007) 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, 1–5 Oct 2007, vol 5, pp 469–473

Dittmar A, Hensch M (2015) Two-level personas for nested design spaces. In: Paper presented at the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Republic of Korea

Drego VL, Dorsey M (2010) The ROI of personas. Forrester Research. https://www.forrester.com/report/The+ROI+Of+Personas/-/E-RES55359

Dursun A, Caber M (2016) Using data mining techniques for profiling profitable hotel customers: an application of RFM analysis. Tour Manag Perspect 18:153–160

Eriksson E, Artman H, Swartling A (2013) The secret life of a persona: when the personal becomes private. In: Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France

Faily S, Flechais I (2011) Persona cases: a technique for grounding personas. In: Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada

Firat AF, Shultz CJ (1997) From segmentation to fragmentation: markets and marketing strategy in the postmodern era. Eur J Mark 31:183–207

Foedermayr EK, Diamantopoulos A (2008) Market segmentation in practice: review of empirical studies, methodological assessment, and agenda for future research. J Strateg Mark 16:223–265

Friess E (2012) Personas and decision making in the design process: an ethnographic case study. In: Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA

Goodwin K, Cooper A (2009) designing for the digital age: how to create human-centered products and services. Wiley, Indianapolis

Gray RM (1984) Vector quantization. IEEE ASSP Mag 1:4–29

Jansen BJ (2009) Understanding user-web interactions via web analytics. Morgan-Claypool lecture series. Morgan-Claypool, San Rafael

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

Jansen BJ, Moore K, Carman S (2013) Evaluating the performance of demographic targeting using gender in keyword advertising. Inf Process Manag 49:286–302

Jansen BJ, An J, Kwak H, Hassen MZ, Cho HY (2016) Efforts towards automatically generating personas in real-time using actual user data. In: Paper presented at the Qatar Foundation Annual Research Conference 2016, Doha, Qatar, 22–23 Mar

Jansen BJ, An J, Kwak H, Salminen JO, Jung SG (2017a) Viewed by too many or viewed too little: using information dissemination for audience segmentation. In: Association for Information Science and Technology Annual Meeting 2017 (ASIST2017), Washington, DC, 27 Oct–1 Nov, pp 189–196

Jansen BJ, Jung SG, Salminen J, An J, Kwak H (2017b) Social analytics data for identifying customer segments for online news media. In: The Third International Workshop on Online Social Networks Technologies, 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA2017), Hammamet, Tunisia, 30 Oct–3 Nov

Jenkinson A (1994) Beyond segmentation. J Target Meas Anal Mark 3:60–72

Jolliffe I (2002) Principal component analysis. Wiley, Hoboken

Judge T, Matthews T, Whittaker S (2012) Comparing collaboration and individual personas for the design and evaluation of collaboration software. In: Paper presented at the Proceedings of the SIGCHI

Conference on Human Factors in Computing Systems, Austin, Texas, USA
Jung S, An J, Kwak H, Ahmad M, Nielsen L, Jansen BJ (2017) Persona generation from aggregated social media data. In: ACM Conference on Human Factors in Computing Systems 2017 (CHI2017), Denver, CO, 6–11 May 2017, pp 1748–1755
Kamboj S, Kumar V, Rahman Z (2017) Social media usage and firm performance: the mediating role of social capital Soc Netw Anal Min 7:51
Kwak H, An J (2014) Understanding news geography and major determinants of global news coverage of disasters. In: Paper presented at the Computation + Journalism Symposium 2014, New York, NY, 24–25 Oct
Kwak H, An J, Jansen BJ (2017) Automatic generation of personas using YouTube Social media data. In: Hawaii International Conference on System Sciences (HICSS-50), Waikoloa, Hawaii, 4–7 Jan 2017, pp 833–842
Lee DD, Seung SH (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
Lin CJ (2007) On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Trans Neural Netw 18:1589–1596
Mao E, Zhang J (2015) 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, 5–8 Jan 2015, pp 3405–3413
Marcus C (1998) A practical yet meaningful approach to customer segmentation. J Consum Mark 15:494–504
Mirzal A (2014) Nonparametric Orthogonal NMF and its Application in Cancer Clustering. In: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Singapore, 2014. Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Springer, Singapore, pp 177–184
Nielsen L, Hansen KS (2014) Personas is applicable: a study on the use of personas in Denmark. In: Paper presented at the Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, Toronto, Ontario, Canada
Odlyzko A (2003) Privacy, economics, and price discrimination on the Internet. In: Paper presented at the Proceedings of the 5th International Conference on Electronic Commerce, Pittsburgh, Pennsylvania, USA
Ortiz-Cordova A, Jansen BJ (2012) Classifying web search queries in order to identify high revenue generating customers. J Am Soc Inform Sci Technol 63:1426–1441
Pruitt J, Adlin T (2005) The persona lifecycle: keeping people in mind throughout product design. Morgan Kaufmann Publishers Inc, Burlington
Pruitt J, Adlin T (2006) The persona lifecycle: keeping people in mind throughout product design. Morgan Kaufmann Publishers Inc, Burlington
Pruitt J, Grudin J (2003) Personas: practice and theory. In: Paper presented at the Proceedings of the 2003 conference on Designing for user experiences, San Francisco, California
Revella A (2015) Buyer personas: how to gain insight into your customer's expectations, align your marketing strategies, and win more business. Wiley, Hoboken
Salminen JO et al (2017) Generating cultural personas from social data: a perspective of Middle Eastern users. In: 2017 5th International Conference on the Future Internet of Things and Cloud Workshops (FiCloudW 2017), Prague, pp 120–125
Shan D, Xu X, Liang T, Ding S (2018) Rank-adaptive non-negative matrix factorization. Cogn Comput 10:506–515
Shapiro BP, Bonoma TV (1984) How to segment industrial markets. https://hbr.org/1984/05/how-to-segment-industrial-markets. Accessed 3 Dec 2017
Shi X, Lu H, He Y, He S (2015a) Community detection in social network with pairwisely constrained symmetric non-negative matrix factorization. In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2015), Paris, France pp 541–546
Shi X, Lu H, He Y, He S (2015b) Community detection in social network with pairwisely constrained symmetric non-negative matrix factorization. In: Paper presented at the Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, Paris, France
Shuradze G, Wagner HT (2016) Towards a conceptualization of data analytics capabilities. In: 2016 49th Hawaii International Conference on System Sciences (HICSS), 5–8 Jan 2016, pp 5052–5064
Smith WR (1956) A product differentiation and market segmentation as alternative marketing strategies. J Advert 21:3–8
Stern BB (1994) A revised communication model for advertising: multiple dimensions of the source, the message, and the recipient. J Advert 23:5–15
Tuna T, Akbas E, Aksoy A, Canbaz MA, Karabiyik U, Aygun BG (2016) User characterization for online social networks. Soc Netw Anal Min 6:104
Xu C (2018) A novel recommendation method based on social network using matrix factorization technique. Inf Process Manag 54:463–474
Zarrinkalam F, Kahani M, Bagheri E (2018) Mining user interests over active topics on social networks. Inf Process Manag 54:339–357
Zhang X, Brown H-F, Shankar A (2016) Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry. In: Paper presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, Santa Clara, California, USA