Capturing the Change in Topical Interests of Personas Over Time

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

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

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

ABSTRACT

In this research, we collect monthly content consumption and demographic data from YouTube over two years for a large media publisher. We use automation to generate 15 personas each month and examine the consistency of the generated personas over time. We find that there are 35 unique personas in total for the entire period, reflecting the changes in the underlying audience population. For each persona, we generate topics of interest and identify the top three monthly topics for each of the 35 personas following an identical algorithmic approach each month. We then compare the sets of topical interests of the personas month-over-month for the entire two-year period. Findings show that there is an average 20.2% change in topical interests and that 68% of the personas experience more topical change than topical consistency. Findings suggest that the topical interests of online audiences are fluid and changes in the underlying audience data can occur within a relatively short period, resulting in the need for constant updating of personas using data-driven methods. The implications for organizations seeking to understand their online audience are that they should employ routine data analysis to detect changes in the audience interests and investigate ways to automate their persona generation processes.

KEYWORDS

Personas; Web Analytics; Online Data Representations

ASIS&T THESAURUS

Information Discovery; Information Behavior; Information and Data Processes

INTRODUCTION

Personas portray segments of audiences, as well as customers or users, as imaginary people (Cooper, 2004), usually within the form of a persona profile containing attributes of the segment that the fictionalized person represents (Nielsen, Hansen, Stage, & Billestrup, 2015). Personas are used widely in a variety of domains for understanding audiences, users, and customers. In this research, we employ personas to represent audience segments for a major online media organization that publishes thousands of news videos yearly.

Personas have been criticized for staling (Chapman & Milham, 2006), meaning that the underlying data from which they were created becomes invalid, resulting in outdated personas. This limitation is especially acute for organizations distributing online content, especially concerning changes in topical interests. It is this last concern, the change in topical interests of personas, that we investigate in this research. Despite the routinely stated critique of personas changing over time (Chapman, Love, Milham, Elrif, & Alford, 2008; Chapman & Milham, 2006; Salminen, Kwak, An, Jung, & Jansen, 2018), we could locate no specific research investigating whether changes in content interests of audience segments, as represented by personas, actually do become outdated. Similarly, there is a lack of research determining how frequently interests do change, if they change at all.

As such, there are several gaps in the literature. *Do topical interests of personas change over time? If so, how often do they change? What is the pace of change? How many topics change? How does one identify when topics change?* These are some of the questions that motivate our research. We could locate no previous empirical analyses of the change in personas' topical interests in any context. We are interested in examining this aspect for personas in pursuit of a larger research project to make personas more data-driven and reflective of audience analytics and segments (Jung, Salminen, An, Kwak, & Jansen, 2018; Salminen, Jansen, An, Kwak, & Jung, 2019).

In this research, we collect user data monthly during a more than two-year period for a large international YouTube content-producing organization with an audience in the hundreds of thousands. To simplify the representation of this large audience population, we generate 15 personas from each monthly dataset following an identical algorithmic methodology for data-driven persona creation (An, Kwak, Salminen, Jung, & Jansen, 2018b). Personas are deployed to represent users or customers in many domains (Friess, 2012). For our research, these generated personas represent the audience segments of the YouTube channel of a media

82nd Annual Meeting of the Association for Information Science & Technology | Melbourne, Australia | 19–23 October, 2019
Author(s) retain copyright, but ASIS&T receives an exclusive publication license
DOI: 10.1002/pra2.00011

organization that currently employs the system used for creating the personas for gaining insights into their audience composition and content preferences.

Associated with the generation of these personas, we identify the top three topical interests of each persona for each month, including the top three interests in the persona profile. We then compare the changes in the topical interests of the personas by month over the entire data collection period. Our findings show that personas' topical interests do change, and these interests can change quite rapidly.

In the following sections, we present a brief literature review, research objectives, methodology, and results. We end with implications and future research directions. See Table 1 for key constructs and definitions applicable to this research.

Persona: An imaginary person created from data that represents a user segment for content, system, or product (Nielsen et al., 2015)
Persona set: The collection of one or more personas that represent the audience population for an organization. In this research, the set of personas is 15, corresponding to the relatively small number recommended in the persona literature (e.g., Goodman, Kuniavsky, & Moed, 2013)
Audience segment: Group of individuals that are similar in specific ways, such as age, gender, interests or behaviors, and usually derived from demographics (e.g., Jenkinson, 1994)
Demographics: Statistical data relating to a population or particular groups within it
Topic: Is the subject of one or more pieces of content. The topic of online content is determined via general processes of topic modeling and text classification (e.g., Xiao, Ji, Li, Zhuang, & Shi, 2018)
Topic modeling: A form of supervised or unsupervised learning where the set of possible topics is inferred from the data (e.g., Hong & Davison, 2010)

Table 1. Key terms employed in this research with definitions.

REVIEW OF LITERATURE

The availability of online content enables users to indicate their preferences for this content via views, likes, comments, and shares (Lee & Tandoc, 2017). With content being produced online, there have been efforts in development services to personalize this content to audiences based on user interests (Karimi, Jannach, & Jugovac, 2018), which requires a focus on topical preferences from audience segments (Li, Bai, Wenjun, & Xihao, 2019; Sánchez & Bellogín, 2019) or on personalization.

Personas are a common method of presenting audience segments (Goodwin & Cooper, 2009); however, there are concerns that multiple or periodic rounds of data collection may be needed to keep the personas updated (Mulder & Yaar, 2007), which can be time consuming and expensive. In fact, without periodic new data collection, end users of the personas are uncertain whether the personas are representative of their *current* target audience. This criticism, widely present in the persona literature (e.g., Chapman & Milham, 2006), is based on the assumption that there can be instability in the audience populations (Drutsa, Gusev, & Serdyukov, 2017) or, as our premise is for this research, in attributes of the current audience. This criticism of staling of the segmentation data assumes that there are periods of instability and change in the underlying populations or data from which the personas are created.

This is an acute concern for online content producers, such as news organizations, social media platforms, news services, and blog sites. The use of online data and metrics allows for the identification of important content for persona generation (Jansen, Jung, Salminen, An, & Kwak, 2017). However, we could find no prior research that would show that the content interests of personas change over time. When using personas for decision making, keeping the personas up-to-date in terms of interests and changes in these interests is of critical importance when creating online content.

For instance, there have been a variety of efforts toward personalization for the delivery of online content based on understanding the interests of audience members. As an example of these efforts, Watters and Wang (2000) extract feature phrases from news content. However, it seems that people prefer a combination of personalization and serendipitous content (Shepherd, Duffy, Watters, & Gugle, 2001), indicating, perhaps, that audiences are interested in new types of content or tire when being exposed to the same topics over time. Therefore, efforts have also focused on modeling audience preferences over time, including efforts based on similarity among audience members (Lv, Meng, & Zhang, 2017) or specific aspects of the content, such as entity identification (Zhang, Boons, & Batista-Navarro, 2019).

Personalizing the delivery of news and other online content depends on the ability to predict user interests, with tailored content being preferred by users, although declared interest and actual interest may differ (Sela, Lavie, Inbar, Oppenheim, & Meyer, 2015). Therefore, developing personalized content systems requires not only detecting and tracking current interests but also predicting what content users would be interested in the future (Mele, Bahrainian, & Crestani, 2019) or at least what related future content (Toraman & Can, 2017) a person might prefer. Although content producing organizations naturally attempt to do this, studies on online news have shown that most news content focuses on consistent themes (i.e., United States, Western Europe), although this is undergoing a change with shifts in news content to Middle Eastern, and Asian countries (Segev, Sheafer, & Shenhav, 2013).

Additionally, prior work (Kwak, An, Salminen, Jung, & Jansen, 2018) has shown that media attention and public attention are both similar and different depending on the resolution of the analysis, with strong regional similarities with both media and public attention but also a substantial number of countries where media attention and public attention are dissimilar by topical interest. In these cases, the public within these countries may be ignored by country-specific content outlets and seek other online sources to address their content needs and desires (Kwak et al., 2018). At the individual level, there has been limited prior work in examining the changes in and the diversity of news content consumption. However, previous research has noted a trend for decreasing content consumption diversity, with the degree of consumption diversity being correlated with the diversity of content available (Zhang, Zheng, & Peng, 2017). However, there is a notable gap in the research in determining the drift or change in the topical interests of an audience for online content.

This is an important research gap to address, as creating personas is not a cheap, easy, or quick process, given it has historically involved ethnography studies or focus groups (Goodwin & Cooper, 2009). When automated methods are employed (An, Kwak, Salminen, Jung, & Jansen, 2018a), there is still effort and time involved in the persona creation process. Therefore, whether or not additional rounds of data collection are needed has practical implications, notably for those organizations with large user populations, such as major online content publishers whose audiences' might fluctuate rapidly.

RESEARCH OBJECTIVES

Our research goals are twofold. First, we investigate whether the topical interests of personas change over time and, if so, how much. Secondly, we determine if the change in topical interest is associated with changes in the set of personas. In pursuit of these goals, we address the following research questions, which are:

- Research Question 01 (RQ01): *Do the topical interests of personas change over time? Is so, how?*
- Research Question 02 (RQ02): *Do the changes in the topical interests of personas vary by the changes in the set of personal? If so, how?*

Our motivation for these questions is to investigate if and how much topical interests of a set of personas change over time to provide insights for organizations concerning how frequently they need to monitor their audiences to keep up-to-date on the interests of audience segments. As audience segmentation is a common practice, we are also interested in whether there are differences in changes in topical interest specifically based on the stability of the audience segmentation, which again provides insight for organizations that use personas or related customer understanding processes.

As such, we define two concepts for this work, which are:

- *Topic Consistency* – the concept of moving from one set of topics to another set of topics within a given period. We address this topic drift by the following two measures, which are:
  - *Consistency* – the persistence in a set of topics from one period to another
  - *Change* – the difference in a set of topics from one period to another
- *Persona Stability* – the concept of the constancy of a set of personas within a given period. We address personas variance by the following two measures, which are:
  - *Stability* – the persistence in a set of personas from one period to another
  - *Instability* – the difference in a set of personas from one period to another

METHODOLOGY

We investigate these questions using data from a major publisher of YouTube content. With the increased availability of online user and customer data, there is the opportunity to use data-driven personas derived directly from a system's users or a company's customer analytics data (Vecchio, Mele, Ndou, & Secundo, 2018), with each persona representing a distinct audience segment. Personas developed from this approach can be created automatically (Jung et al., 2017), which we did for this research.

Data Collection

This organization had more than 665,404 subscribers and more than 4,800 of pieces of online content at the time of the study. As such, the organization is representative of those publishers that distribute content via major online platforms, such as other content producers, app developers, software-as-a-service (SaaS) providers, and other media outlets. These organizations typically have a large and potentially diverse audience, varied online content, and the potential for changes in the underlying audience population.

The YouTube Analytics API provides statistics for each piece of content and various user profile data, (e.g., gender, age, country location) at an aggregate level. Individual user data is not provided to safeguard users' privacy. Via the YouTube Analytics API, we collect the detailed record of content interactions by country, gender, and age group for each piece of content. We collect data for all videos of the channel published each month from October 2016 through January 2019. There was 1 month where the data collection failed, so we have 26 months of data available for this research.

System for Generating Personas

In terms of producing the personas, the applied approach for generating data-driven personas is discussed in prior work (An et al., 2018a); therefore, we only briefly present it here. The approach relies on non-negative matrix factorization (NMF) (Lee & Seung, 1999) to take the aggregated user data

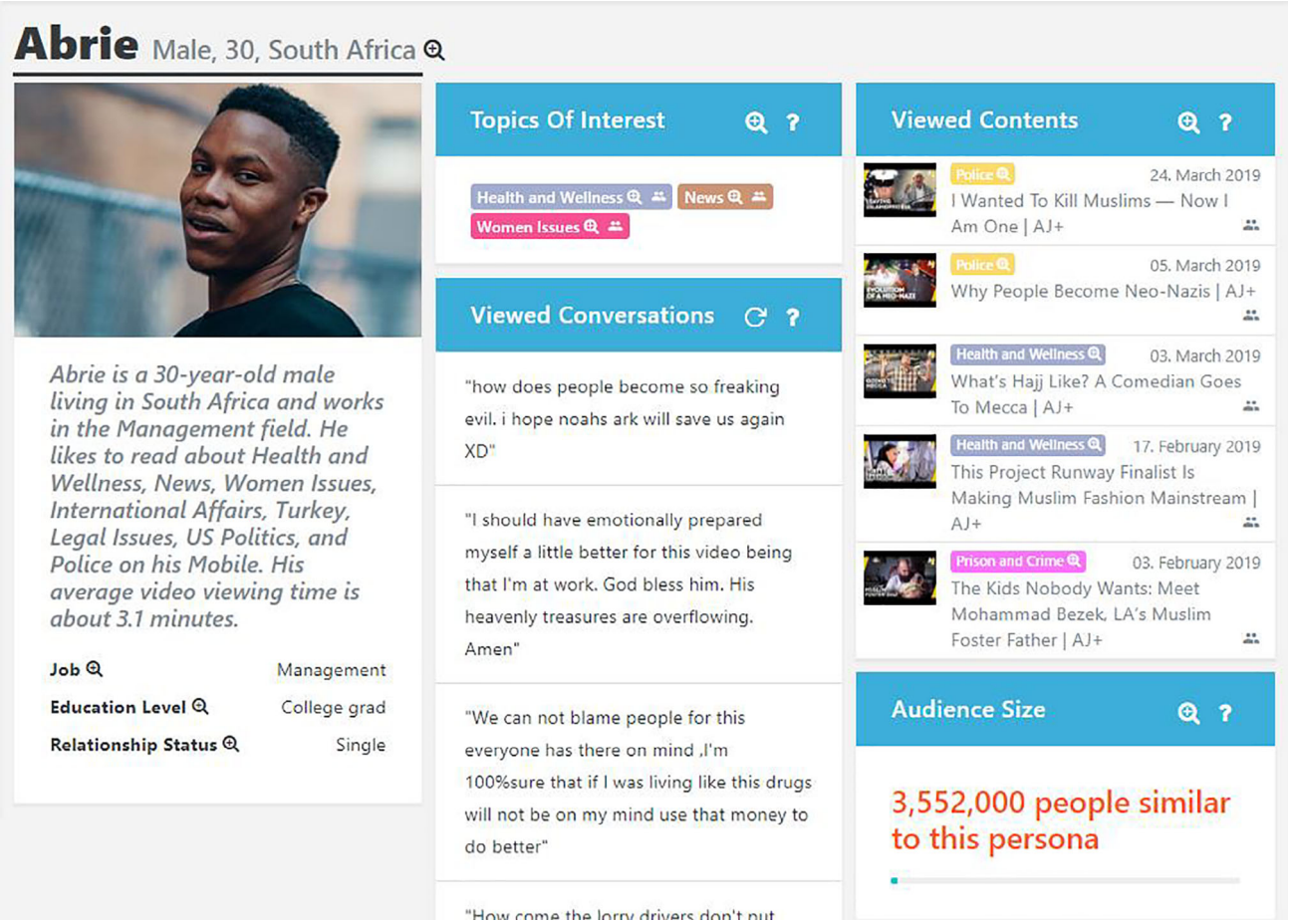

Figure 1. Example of a data-driven persona profile used in this research. The persona is created automatically from audience demographic and associated behaviors data of the organization's online content. One component of the persona profile is the Topics of Interest, which is the focus of this research. Topics of Interest are computed using the Z-score over the persona set.

and identify unique behavioral patterns (Sánchez & Bellogín, 2019), associating these unique sets of behavioral patterns, with demographic attributes, and then using other algorithmic approaches to generate a complete persona profile.

In our example, we apply NMF to the video content views retrieved from You Analytics for the organization, resulting in a set of unique view patterns. Also, using data from YouTube Analytics, each unique behavior pattern is associated with one or more set of demographics. The demographic with the highest NMF weight is associated with the behavior pattern. These are the base personas that the system then enriches with name, photo, etc. Each video is automatically topically classified, providing topics of interests for each persona. An example of a complete persona profile from the system is shown in Figure 1. This approach has several advantages, including that is responsive to interactions with both existing and new content, which is important as our datasets are cascading (i.e., existing content will get new user interactions, and new content is added that has no prior user interactions). The result of the process is the set of the most distinct personas in terms of both behavioral and demographic user attributes.

We apply the identical methodological approach to each monthly dataset, generating 15 personas each iteration. Although 15 personas are three to five times the standard number of personas (Pruitt & Grudin, 2003), we deem the higher number of personas reasonable for organizations with varied online audiences. The result of monthly data collection and repeated analysis is a series of monthly sets of organizational personas over the period. Once we have the complete series of 26 data sets, we then present the 15 top personas for each month. We use a system to automate this process and display the listings of personas, as shown in Figure 2 (i.e., a listing of multiple personas) and Figure 3 (i.e., one persona listing for readability of the topical interests).

Again, the topic classification method of automatic persona generation is explained in prior work (An et al., 2018a). However, as a brief overview, one of the benefits of using NMF for generating these personas, representing audience segments, is a clear association between the audience segments' interest and non-interest in specific digital content.

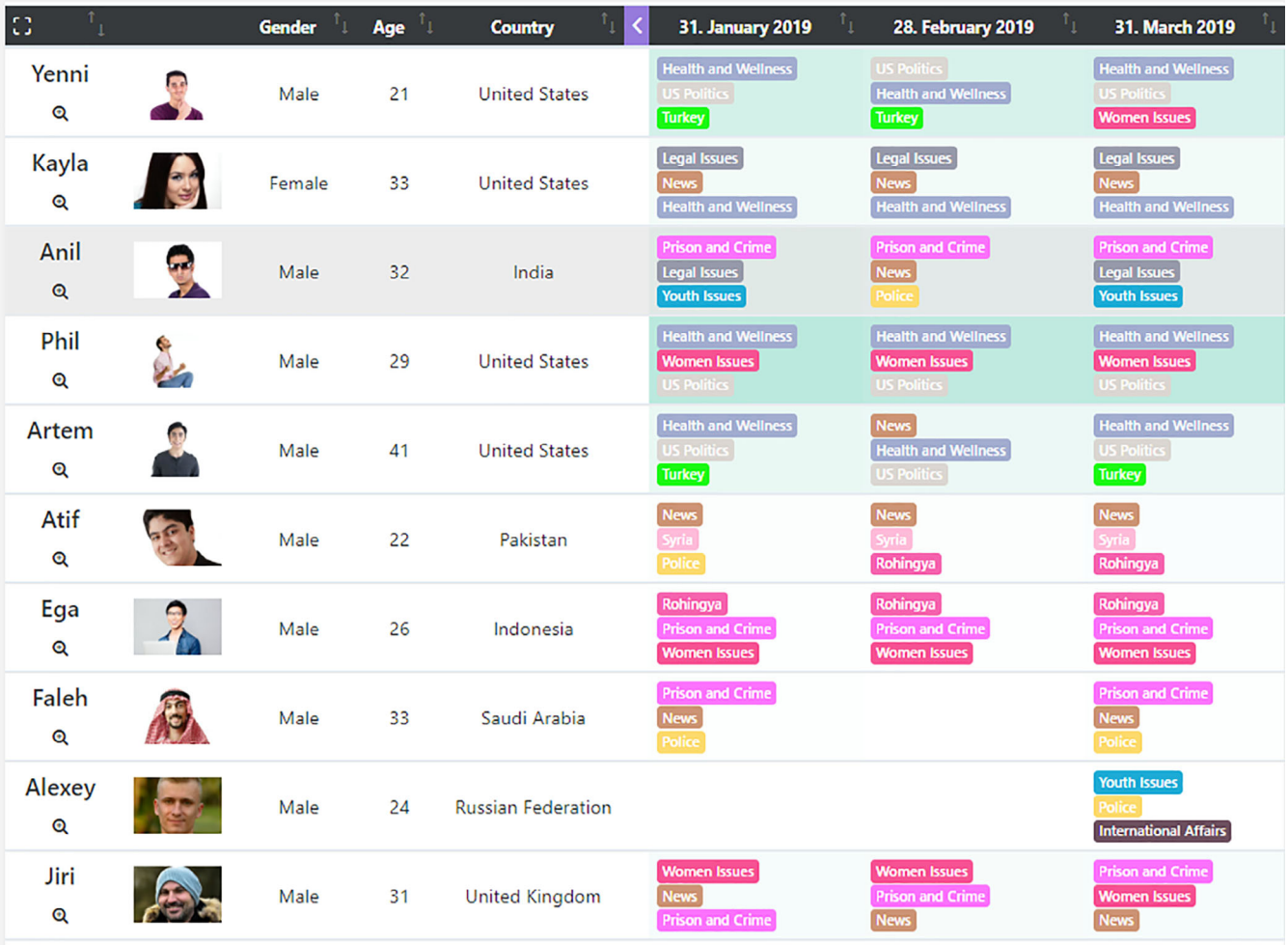

Figure 2. Screenshot of a portion of the persona sets. The left lists the specific personas. The top displays the month of the data collection. Along the right is the listing the top three topics of interest for each persona for each monthly data collection.

Figure 3. Screenshot of one persona with monthly data collection date and specific topics of interest for each period.

Beginning with this association, we can identify content 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 audience segment. To compute the similarity of content in a robust way, we explicitly express the content features and how we leverage them for persona creation.

We define a matrix that captures the features of each piece of content. We then derive another matrix that represents an association between an audience segment and content features. Then, using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), we cluster content into topical categories to which we assign a human interpretable label.

We quantify the topical interests for the personas by computing the Z-score for F_t^i. For the topic t and the persona i, as follows:

$$F_t^i = \frac{F_t^i - avg(F_t)}{\sigma} \quad (1)$$

where F_t is a set of F_t^i for any existing persona X, and σ is the standard deviation of the F_t. A higher Z-score means that a persona is more likely to view content belonging to a topic than the other personas. The result is a ranking of topics for each persona, as shown in Figure 4. The bars to the left

indicate topics of most interest and the bars to the right indicate topics of least interest, relative to the entire set of personas generated for that monthly data collection.

In summary, we used the identical algorithmic process to generate both the personas and the topical interests of these personas for each monthly data collection. As such, the results from an identical process allowed for the comparison of topical interests across all months for each persona, which is the focus of the research we report here.

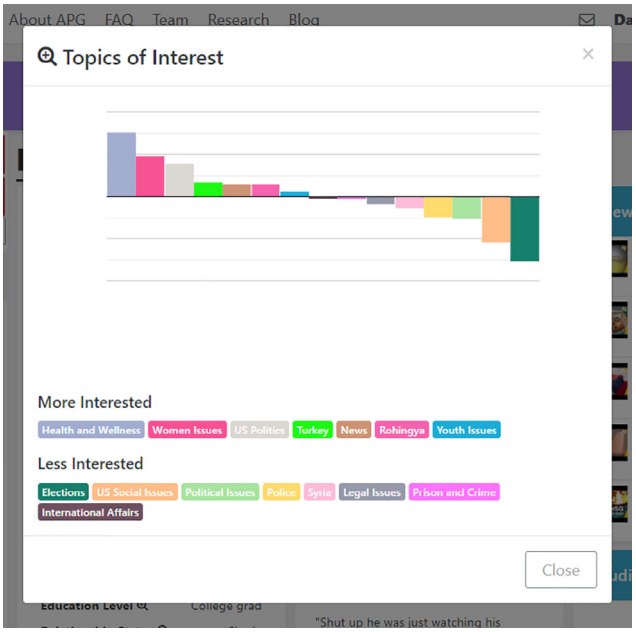

Figure 4. Screenshot of the ranking of topics by the interest for a specific persona. A similar calculation is computed for each persona for each data collection period. Note that a lower ranked topic does not imply non-interest, just less interest than a higher ranked topic.

Methodology for Comparing Topical Interest

We compare the top three topical interest of each persona month-over-month (MoM) (i.e., overlap expressed for the previous month. We selected the top three topics as this is what is presented in the persona profiles and most likely to be used by the organization for audience analysis. We did not consider the order of the three topics. Specifically, we define our consistency metric as the following:

$$\text{MoM} = (T_{PM} \Pi T_{CM}) / T_{PM} \quad (2)$$

Where T_{PM} is the set of persona's topics in the prior month, T_{CM} is the set of persona's topics in the current month. Once we have the overlap for this metric, it is trivial to calculate the MoM change (i.e., the MoM change is one minus the MoM consistency value). Returning to our research questions, we now examine the change in topics of personas over time.

RESULTS AND DISCUSSION

Returning to our research questions, we first present overall findings and then examine RQ01 (*Do the topical interests of personas change over time?*) and RQ02 (*Do the changes in the topical interests of personas vary by the changes in the persona set?*).

Exploratory Results

Table 2 shows the aggregate distribution of topics during the 26-months of data collection. As shown in Table 2, the topics were stratified into three general clusters, a highly popular cluster (General News, Prison, and Crime, etc.), a mid-popular cluster (Women Issues, Syria, etc.), and less popular cluster (Turkey, US Politics, etc.).

	Topic	Count	%
1	General news	218	18.7
2	Prison and crime	154	13.3
3	Health and wellness	142	12.1
4	Police and law enforcement	129	11.0
5	Women issues	104	8.9
6	Syria and the syrian conflict	67	5.7
7	Youth issues	64	5.5
8	Legal issues	59	5.0
9	Refugee issues, Rohingya	59	5.0
10	Turkey MENA relations	49	4.2
11	US politics	49	4.2
12	International affairs	48	4.1
13	Elections (world)	12	1.0
14	US social issues	11	0.9
15	Political issues	5	0.4
		1170	100

Table 2. The distribution of topics over the entire data collection period of 26 months for 15 personas in each month.

Month-over-Month Change in Personas

To examine RQ01, we computed the MoM change, with findings presented in Table 3. As shown in Table 3, the average MoM change was 29.2% of topics among each set of 15 personas per month (standard deviation of 27.2%). The minimum MoM change was 0.0%, and the maximum MoM change was 82.6%. The median was 25.0%.

	MoM consistency (%)	MoM change (%)
Average	70.8	29.2
Std. dev.	72.8	27.2
Max	17.4	82.6
Min	100.0	0.0
Median	75.0	25.0

Table 3. Overall statistics for the topic of interest consistency and changes for the entire set of personas over the entire data collection period.

Table 4 shows the number of personas that experienced primarily higher months of change, higher months of consistency, and with change and consistency the same. As shown in Table 4, more than 68% of the personas experience a higher level of MoM change compared to consistency during the data collection period, and 5.7% of the personas experienced an equal amount of change.

So, returning to our RQ01, the topical interests of persona do change, and these interests can change rather rapidly. The topical interests by persona varied by approximately 29% MoM and more than 68% of the personas experienced more topical change than consistency MoM during the data collection period.

MoM	Number of Personas	%
Greater change	24	68.6
Equal	2	5.7
Greater consistency	9	25.7
	35	100.0

Table 4. MoM change and consistency of number of personas.

Change in Personas During the Entire Period

Moving to RQ02, concerning if the change in topical interest is correlated with the change in personas set, we examine the persona ordered by stability.

Table 5 shows the persistence of the personas' appearance over the period. As shown, a set of personas can be defined as: core audience, as they are always or nearly always in the audience, that are loyal customers (stability of 19–26 months); a set of customers that are generally but not always in the audience (8–12 months); and then a set of customers that stay in the audience for a short time (1–7 months).

We ran a Spearman correlation test to determine the strength and direction of the monotonic relationship between personas stability and topical consistency. Results of the test indicate that there is a significant and positive association between persona stability and topic consistency ($rs(27) = .73$, $p < .05$). Figure 5 shows the relationship between persona stability and topical consistency. As shown, there is a positive relationship but also with outliers and some variance. So, the topical interests are affected by the stability of the persona set. The more consistent a persona is with an organization the more stable is their top interests.

Verification of Organizational Content

One factor that could have impacted the topical analysis is the content produced by the organization could have changed in a given month relative to what was published previously. To investigate this possibility, we compared the aggregated consistency of topics month-over-month for all content. We further compute the Jaccard coefficient that measures the

Months	Percentage of period (%)	Number of personas	% of personas
26	100	5	13.17
24	92	3	7.89
21	81	1	2.64
20	77	1	2.63
19	73	1	2.63
12	46	1	2.63
11	42	1	2.63
10	38	1	2.63
9	35	2	5.26
8	31	3	7.89
7	27	4	10.53
6	23	2	5.26
3	12	1	2.63
2	8	1	2.63
1	4	11	28.95
		38	100.00

Table 5. The stability of personas during the data collection period. Months is the number of months that a given persona appeared in the data set (i.e., the measure of stability).

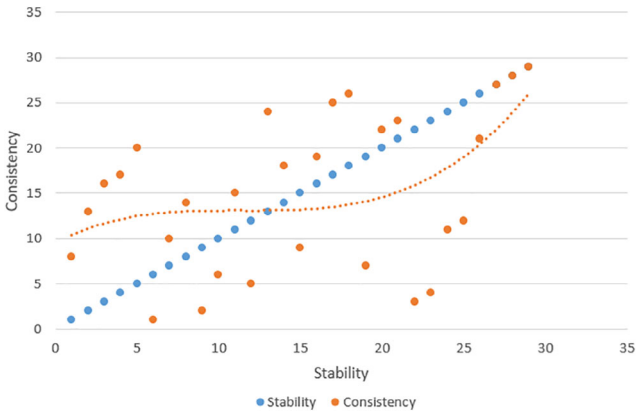

Figure 5. Plot with trend line and the equation of persona stability (x-axis) and topical consistency (y-axis).

overlap of two sets, which in our case are *set 1* = the topics in the prior month; and *set 2* = the topics in the current month. We use the first month as the starting baseline. Equation 3 shows the calculation for the Jaccard coefficient.

$$J = \frac{N_c}{N_a + N_b - N_c} \quad (3)$$

where:

N_a = Number of topics in the prior month.

N_b = Number of topics in the current month.

N_c = Number of topics in the intersection of the Prior Month and the current month.

As shown in Table 6, the average consistency was 97.0% (3.1% change). The average Jaccard coefficient was 0.93, indicating a high level of similarity among the topics published. These scores indicate further that the changes in interests noted in the topical drift of the personas were the result of changes in audience, changes in the audience preferences, or changes in both and not due to the availability of or changes in the content from the organization.

DISCUSSION AND IMPLICATIONS

In this research, we show that the topical interests of personas do change over time, confirming a central thesis in the persona literature. A key implication is that organizations using personas should engage in continuous data collection to ensure that potential changes in their audience base and interests are reflected in the personas they are using.

Month	Matching topics	Missing topics	Total topics	Consistency with prior month (%)	Jaccard coefficient with prior month
2	14	0	14	100.0	0.93
3	12	1	13	92.3	0.80
4	12	0	12	100.0	0.92
5	11	3	14	78.6	0.73
6	14	1	15	93.3	0.93
7	13	0	13	100.0	0.87
8	13	0	13	100.0	1.00
9	13	0	13	100.0	1.00
10	12	0	12	100.0	0.92
11	11	1	12	91.7	0.85
12	12	1	13	92.3	0.92
13	13	0	13	100.0	1.00
14	12	0	12	100.0	0.92
15	12	0	12	100.0	1.00
16	12	0	12	100.0	1.00
17	12	0	12	100.0	1.00
18	12	0	12	100.0	1.00
19	12	0	12	100.0	1.00
20	12	1	13	92.3	0.92
21	12	0	12	100.0	0.92
22	10	0	10	100.0	0.83
23	10	2	12	83.3	0.83
24	12	0	12	100.0	1.00
25	11	0	11	100.0	0.92
26	11	0	11	100.0	1.00
Total	300	10			
Average			12.4	97.0% (3.0% change)	0.93

Table 6. Comparison of the overall topics published by the organization MoM (with month 1 as the base). Matching topics is the number of topics in the current month that was also in the prior month. Total topics are the number of topics that occurred among all personas for the current month. Consistency is the MoM percentage of topics in the current month relative to the prior month.

Regarding online content consumption, our findings confirm and expand those reported in (Zhang et al., 2017) that noted a decrease in the consumption of content diversity over time (i.e., the content people consumed tended toward sameness over time).The repeated data collection can be done via traditional means (i.e., surveys, analytics) or using automated methods, as done here. The advantage of using online data sources and algorithmic approaches for persona creation is the effortless comparison of changes over time and the use of identical methods for persona generation.

Our research also addresses the widely cited criticism of Chapman and Milham (2006) that argues *personas are beyond the scope of scientific validation* because their creative process cannot be replicated. By quantifying and standardizing the persona generation process and the attributes of the generated personas (in this case, their topics of interest), data-driven personas can be brought into the realm of scientific inquiry. While Chapman and Milham (2006) attempted this years ago, there has been little considerable progress since then regarding the longitudinal change of personas, a predisposition taken for granted in the literature. In this research, we clearly show how to quantify and measure this change (consistency), finding results that validate the need for periodic updating of personas or any other customer/user profiles that are based on evolving analytics data. This novel aspect of this research, the concept of analyzing

the change of audience behaviors as personas over time, is an innovative approach only made practically possible by the availability of large-scale online data that can be used for developing behavioral personas, as noted in prior work (e.g., Zhang, Brown, & Shankar, 2016).

Although there have been discussions in prior works concerning the possible change in personas (Chapman et al., 2008; Chapman & Milham, 2006), there is a subtle distinction in our research reported here, which concerns a change in content preferences, which we show is due to both (a) an underlying change in the audience and (b) some of the current audience undergoing changing preferences. For online content producing organizations, it highlights the need to monitor both large demographical changes in the audience and the more fluid fluctuations in audience preferences that have an impact on content consumption.

Future research could investigate the pace of topic change for a high-frequency content producer and a low-frequency content producer. The assumption is that the personas (i.e., audience composition) remain more stable for low-frequency content producers because no new audiences are attracted by the changing topic profile of the content. However, this assumption needs investigation.

Other future research could entail using the presented comparison approach to determine the optimal number of topical interests over a period of time, which could be the number where the period of change is minimized. This also highlights a possible limitation in our research, in that we might overestimate topic change by limiting our focus to only the top three topics. Future research could address this limitation. Finally, we could also examine the effect of detecting topical shifts on content automatically (Alkhodair, Ding, Fung, & Liu, 2019) for integration into an overall data-driven persona system. This could assist content companies in avoiding content becoming stale or having audience segments become bored with the content that is too similar to what they have consumed in the past. Online content companies could anticipate new content that would excite the audience via new content for audience segments.

CONCLUSION

We confirm that topical interests of online audiences can change over time, empirically validating a criticism of personas that their updating requires continual data collection and confirming that the underlying audience population can change. In the data set used here, there were substantial changes in topics, even within a relatively short period. This indicates that organizations incorporating personas or seeking to understand their audiences, in whatever manner the data is collected, should frequently update their audience insights, as changes are likely. Naturally, some of the answers to our research questions may vary organization to organization.

REFERENCES

Alkhodair, S. A., Ding, S. H. H., Fung, B. C. M., & Liu, J. (2019). Detecting breaking news rumors of emerging topics in social media. *Information Processing & Management.*

An, J., Kwak, H., Salminen, J., Jung, S. G., & Jansen, B. J. (2018a). Customer segmentation using online platforms: Isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, *8*(1), 54.

An, J., Kwak, H., Salminen, J., Jung, S. G., & Jansen, B. J. (2018b). Imaginary people representing real numbers: Generating personas from online social media data. *ACM Transactions on the Web*, *12*(4), 27.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Drichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chapman, C. N., Love, E., Milham, R. P., Elrif, P., & Alford, J. L. (2008). *Quantitative evaluation of personas as information*. In paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Chapman, C. N., & Milham, R. P. (2006). *The personas' new clothes: Methodological and practical arguments against a popular method*. In paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

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

Drutsa, A., Gusev, G., & Serdyukov, P. (2017). Periodicity in user engagement with a search engine and its application to online controlled experiments. *ACM Transactions on the Web*, *11*(2), 1–35.

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, TX.

Goodman, E., Kuniavsky, M., & Moed, A. (2013). *Observing the user experience: A practitioner's guide to user research* (2nd ed.). Morgan Kaufmann.

Goodwin, K., & Cooper, A. (2009). *Designing for the digital age: How to create human-centered products and services*. Indianapolis, IN: Wiley.

Hong, L., & Davison, B. D. (2010). *Empirical study of topic modeling in Twitter*. In paper presented at the Proceedings of the First Workshop on Social Media Analytics, Washington, DC.

Jansen, B. J., Jung, S.-G., Salminen, J., An, J., & Kwak, H. (2017). Viewed by too many or viewed too little: Using information dissemination for audience segmentation. *Proceedings of the Association for Information Science and Technology*, *54*(1), 189–196.

Jenkinson, A. (1994). Beyond segmentation. *Journal of Targeting, Measurement and Analysis for Marketing*, *3*(1), 60–72.

Jung, S., An, J., Kwak, H., Ahmad, M., Nielsen, L., & Jansen, B. J. (2017). *Persona Generation from Aggregated Social Media Data*. In paper presented at the ACM Conference on Human Factors in Computing Systems 2017 (CHI2017), Denver, CO.

Jung, S., Salminen, J., An, J., Kwak, H., & Jansen, B. J. (2018). *Automatically conceptualizing social media analytics data via personas*. In paper presented at the The International AAAI Conference on Web and Social Media (ICWSM 2018), San Francisco, CA.

Karimi, M., Jannach, D., & Jugovac, M. (2018). News recommender systems – Survey and roads ahead. *Information Processing & Management*, *54*(6), 1203–1227.

Kwak, H., An, J., Salminen, J., Jung, S.-G., & Jansen, B. J. (2018). *What we read, what we search: Media attention and public attention among 193 countries*. In paper presented at the Proceedings of the 2018 World Wide Web Conference, Lyon, France.

Lee, D. D., & Seung, S. H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Lee, E.-J., & Tandoc, E. C. (2017). When news meets the audience: How audience feedback online affects news production and consumption. *Human Communication Research*, *43*(4), 436–449.

Li, C., Bai, J., Wenjun, Z., & Xihao, Y. (2019). Community detection using hierarchical clustering based on edge-weighted similarity in cloud environment. *Information Processing & Management*, *56*(1), 91–109.

Lv, P., Meng, X., & Zhang, Y. (2017). FeRe: Exploiting influence of multi-dimensional features resided in news domain for recommendation. *Information Processing & Management*, *53*(5), 1215–1241.

Mele, I., Bahrainian, S. A., & Crestani, F. (2019). Event mining and timeliness analysis from heterogeneous news streams. *Information Processing & Management*, *56*(3), 969–993.

Mulder, S., & Yaar, Z. (2007). *A practical guide to creating and using personas for the web*. New Riders.

Nielsen, L., Hansen, K. S., Stage, J., & Billestrup, J. (2015). A template for design personas: Analysis of 47 persona descriptions from danish industries and organizations. *International Journal of Sociotechnology and Knowledge Development*, *7*(1), 45–61.

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, CA.

Salminen, J., Jansen, B. J., An, J., Kwak, H., & Jung, S. G. (2019). Automatic persona generation for online content creators: Conceptual rationale and a research agenda. In L. Nielsen (Ed.), *Personas - User focused design* (pp. 135–160).

Salminen, J., Kwak, H., An, J., Jung, S. G., & Jansen, B. J. (2018). Are personas done? Evaluating their usefulness in the age of digital analytics. *Persona Studies*, *4*(2), 47–65.

Sánchez, P., & Bellogín, A. (2019). Building user profiles based on sequences for content and collaborative filtering. *Information Processing & Management*, *56*(1), 192–211.

Segev, E., Sheafer, T., & Shenhav, S. R. (2013). Is the world getting flatter? A new method for examining structural trends in the news. *Journal of the American Society for Information Science and Technology*, *64*(12), 2537–2547.

Sela, M., Lavie, T., Inbar, O., Oppenheim, I., & Meyer, J. (2015). Personalizing news content: An experimental study. *Journal of the Association for Information Science and Technology*, *66*(1), 1–12.

Shepherd, M., Duffy, J. F., Watters, C., & Gugle, N. (2001). The role of user profiles for news filtering. *Journal of the American Society for Information Science and Technology*, *52*(2), 149–160.

Toraman, C., & Can, F. (2017). Discovering story chains: A framework based on zigzagged search and news actors. *Journal of the Association for Information Science and Technology*, *68*(12), 2795–2808. https://doi.org/10.1002/asi.23885

Vecchio, P. D., Mele, G., Ndou, V., & Secundo, G. (2018). Creating value from social big data: Implications for smart tourism destinations. *Information Processing & Management*, *54*(5), 847–860.

Watters, C., & Wang, H. (2000). Rating news documents for similarity. *Journal of the American Society for Information Science*, *51*(9), 793–804.

Xiao, D., Ji, Y., Li, Y., Zhuang, F., & Shi, C. (2018). Coupled matrix factorization and topic modeling for aspect mining. *Information Processing & Management*, *54*(6), 861–873.

Zhang, H., Boons, F., & Batista-Navarro, R. (2019). Whose story is it anyway? Automatic extraction of accounts from news articles. *Information Processing & Management*.

Zhang, L., Zheng, L., & Peng, T.-Q. (2017). Structurally embedded news consumption on mobile news applications. *Information Processing & Management*, *53*(5), 1242–1253.

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, San Jose, CA.