2016 4th International Conference on Future Internet of Things and Cloud Workshops

Towards Automatic Persona Generation Using Social Media

Jisun An, Hoyoun Cho, Haewoon Kwak, Bernard J. Jansen
Qatar Computing Research Institute
Hamad bin Khalifa University
Doha, Qatar
jan@qf.org.qa, monohoyoun@gmail.com,
hkwak@qf.org.qa, jjansen@acm.org

Mohammed Ziyaad Hassen
Strategy & Development Division
Al Jazeera
Doha, Qatar
HassenM@aljazeera.net

***Abstract*— The use of personas is an interactive design technique with considerable potential for product and content development. However, personas have typically been viewed as fairly static. In this research, we implement an approach for creating personas in real time, based on automated analysis of actual social media data, integrating data from Facebook, Twitter, and YouTube channels for a large commercial organization. From Twitter, we gather user insights representing interests and viewpoints, leveraging approximately 195,000 follower profiles. From YouTube, we gather demographic data and topical interests, leveraging more than 188,000 subscriber profiles and millions of user interactions. From Facebook, we collect instances of hundreds of thousands of link sharing by more than 54,000 social media followers, specifically examining the domains these users share. We integrate the social media data from all three platforms in order to demonstrating that this data can be used to develop personas in real-time. The research results provide insights into competitive marketing, topical interests, and preferred system features for the users of the online news medium. Research implications are that personas can be generated in real-time, instead of being the result of a laborious, time-consuming development process.**

Keywords*— Persona; marketing; online news; design method; scenario; user-centered design.

I. INTRODUCTION

Used in marketing and advertising for abstract user representation, the concept of personas has spread to a variety of other fields, such as system design [1]. A persona is a representation of a group or segment of users, sharing common behavioral characteristics. Although representing a segment of users, a persona is generally developed in the form of an explicit but fictitious individual, accompanied by a detailed narrative that represents the collection of users possessing similar behaviors or characteristics.

In order to make the fictitious individual appear as real person to the content or product developers, the persona narrative can contain a variety of both demographic and behavioral details about socio-economic status, gender, hobbies, family members, friends, possessions, among many other data. Also, the narrative of a persona can addresses the goals, needs, wants, frustrations, and other emotional aspects of the fictitious individual that are pertinent to the content or product being designed. The persona is usually given a name and an image to assist the product developers in focusing on the particular user segment [2].

II. RELATED WORK

Personas are beneficial for guiding development decisions and evaluating project ideas, from systems to services to advertisements. Within the domain of system design, personas are considered and have been shown to be a worthwhile technique in directing designers and developers. A persona is created for clarifying and synthesizing descriptions of user segments, with the idea that a persona assists a designer in focusing on the behavior patterns, wants, and needs of the segment of users. Systems, especially those deployed to market, can have multiple user segments that require multiple personas, so the development of the granularity of data and execution of data integration in a meaningful way can be complex.

Ideally, the construction of a persona is based on actual field data research of a product's intended user base(s). Typically, this has been a major issue, as user data has generally been gathered via surveys, focus groups, or ethnography methodologies. However, a common problem with creating personas via these methods is that they are many times not based on first-hand user data [3] or the data set is not of a sample size that can be considered statistically significant. Additionally, utilizing these processes for persona development can be costly and time-consuming [4].

As a result, in practice, persona development does not come from user studies or actual current user behavioral data, but the personas are based on the assumptions, experiences, or expectations of executives, marketers, or designers or the personas quickly become out of date. This results in personas that are not believable or do not actually represent the real current or targeted users [5, 6]. The problem is exacerbated in actual commercial products in fast moving and competitive market areas, where the user base is multivariate and always open to the possibility of flux.

Therefore, there are numerous unanswered questions concerning the automatic generate of personas. *Can the usage of actual user behavioral data be used for personas development? Can online data provide the rich demographic insights common in personas? Can personas be developed in*

978-1-5090-3946-3/16 $31.00 © 2016 IEEE
DOI 10.1109/W-FiCloud.2016.51

near real time? Can personas be continual updated? These are some of the questions that motivate our research.

III. Research Objective

In this research, we propose that real user behavior and related demographic data concerning users of a product, service, or content [7] can be rapidly and inexpensively collected from variety of social platforms and analyzed in order to generate personas in real-time. Achieving such an objective means that personas (1) represent the current users of the product and (2) are sensitive that usage can dramatically change based on audience interests or shift over time. This is the research objective that we investigate, and we present the current state of research and system development here.

IV. Data Collection and Research Design

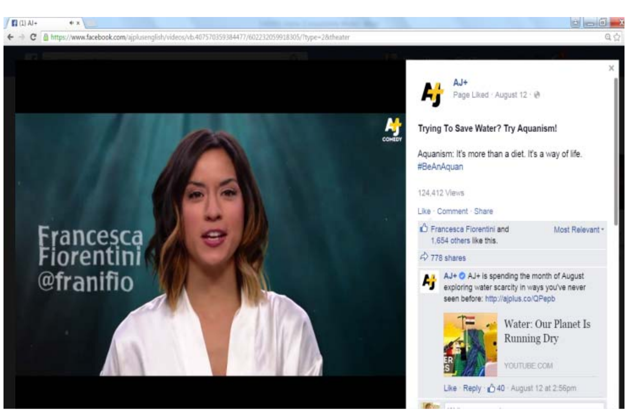
We validate our premise using actual user data from AJ+, an online media and mobile outlet. In the news industry, audience preferences have been somewhat ignored by journalists mainly because of the lack of accurate measurements before the era of online news. In fact, several studies point out large differences between news production and consumption patterns by using the number of articles on particular topics and their views in news websites [8]. Correct understanding of audiences becomes more important to attract audience and increase the consumption of digital content.

Therefore, we consider AJ+ an excellent organization for both data collection, and our research with actual AJ+ user data shows the value of automatic persona generation for news readership research in particular, although we consider the results transferable to other industry segments.

A. Data Collection Organization

AJ+ (http://ajplus.net/) is an online news channel from Al Jazeera Media Network that is natively digital with a presence only on social media platforms (see **Error! Reference source not found.** for an example) and on a mobile app. Its media concept is unique in that AJ+ was designed from the ground up to serve news in the medium of viewer, versus a teaser redirecting to a website.

Fig. 1. Example of AJ+ Facebook Video with captions, Likes, Shares, Comments, and Replies.



AJ+ is based primarily on social platforms but also has a presence on iOS and Android apps. Therefore, the digital content is specifically designed to be viewed in the Facebook newsfeed, YouTube Channel, or Twitter Timeline for the audience segments that are most active on those platforms.

AJ+ has been innovative in its experimentation with storytelling formats, app design, and video development, receiving significant press [9]. At the time of this study, AJ+ was the second largest producer of video on Facebook, had more than 3 million Facebook followers, 195,000 Twitter followers, and 188,000 YouTube subscribers. AJ+'s engagement rate at the time of the study was 600% (i.e., their news products reach and are engaged by 6x their follower base).

Given the prerequisite for rapid and media specific development of content in a competitive and fluid information market [10], AJ+ has a critical need for automatic and real time generation of personas to guide digital content, media, and system planning and design.

Therefore, in pursuit of our overall research objective of automatically generating personas in real time, for the research reported in this manuscript, we are specifically interested in understanding the AJ+ audience by identifying (1) whom are they reaching (i.e., market segment) and (2) what competitive (i.e., non-AJ+) content are associated with each market segment.

From this, combined with other user data [11, 12], we can validate the design a system using actual user data for personas generation in near real time.

B. Data Collection (Facebook)

For data collection, we primarily, for this portion of the research, focus on the Facebook channel, due to space limitations, although we do include Twitter and YouTube features in the system development and present a brief overview of these data integrations.

Facebook is a major driver of AJ+ viewers. Even though Facebook itself has a rich collection of user data — the information typically considered as factors of personas, such as socio-economic status, gender, hobbies, family and friends — but it is not open to AJ+ (or any other Facebook applications) due to the tight privacy controls of many users in Facebook. Even a list of liked items of a user is not publicly accessible by the Facebook API, without an user's explicit permission. Thus, for this portion of the research, we examined the publicly available data of AJ+ Facebook users, specifically the linked content (mostly non-AJ+ links) they shared in their timelines (see **Error! Reference source not found.**).

A set of shared URLs of a user can become a good indication of what digital content the user is interested in and is an excellent source of competitive intelligence for an organization such as AJ+.

Beginning with the AJ+ Facebook page, we extract a list of users who liked any of AJ+ posts. For those who liked AJ+ posts, we then collected all URLs they shared in their timelines. We note that this set of URLs includes non-AJ+ content.

Fig. 2. An example of a URL link shared in user's timeline. Note: Time.com, not AJ+ link

Thus, the dataset shows what users are interested in beyond AJ+ content. For example, if a user shares URLs to imdb.com and netflix.com, then one can tell the user is interested in movies. This detailed user preference cannot be captured by simple demographics offered by Facebook Insights. We then set out to determine if we could cluster users by the domains that they shared.

TABLE I. MOST FREQUENTLY SHARED DOMAINS BY AJ+ USER USERS (DOMAINS WITH MORE THAN 30 THOUSAND SHARES) IN OUR DATASET.

Ranks	Domain	Shares	Percentage
1	youtube.com	1,784,299	70.24%
2	huffingtonpost.com	156,113	6.15%
3	theguardian.com	64,827	2.55%
4	soundcloud.com	60,854	2.40%
5	nytimes.com	60,630	2.39%
6	rt.com	59,997	2.36%
7	wordpress.com	56,628	2.23%
8	buzzfeed.com	48,958	1.93%
9	yahoo.com	39,677	1.56%
10	vimeo.com	38,060	1.50%
11	washingtonpost.com	37,296	1.47%
12	cnn.com	34,765	1.37%
13	upworthy.com	33,623	1.32%
14	ask.fm	32,368	1.27%
15	dailymail.co.uk	32,191	1.27%

As a result of data collection, from 2,785 AJ+ posts, we identified 8,065,350 shared links by 54,892 users. Table 1 shows the aggregate domain analysis (top fifteen domains), offering interesting insights about the AJ+ user base. Aside from the comprehensive YouTube domain, there are sites for traditional news (e.g., cnn.com), foreign news (e.g., rt.com), alternative news (e.g., upworthy.com) entertainment (e.g., soundcloud.com), social question and answering (e.g., ask.fm), and political action (e.g., thefreethoughtproject.com).

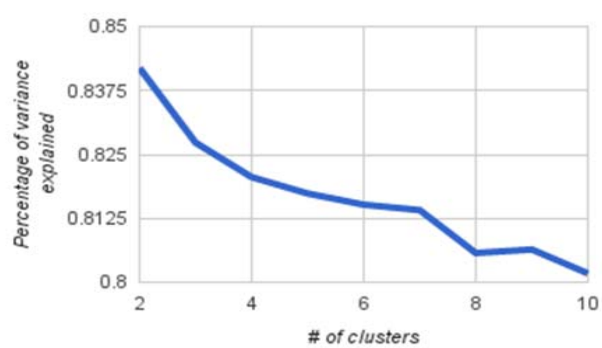
V. DATA ANALYSIS (FACEBOOK)

To automatically generate personas from the social media data, we firstly find a set of users who share the similar behavioral patterns, then, we identify distinct demographic characteristics of those users to create the personas.

As a basis for personas generation, we first cluster users based on behavior data, specifically for Facebook based on their shared links as surrgates for shared content. We specifically use top-level domains of the links; we consider a "domain" as a "word". We then conduct term-frequency – inverse document frequency (tf-idf) vectorization to differently weigh shared domains according to their importance. We remove outliers of less than 5 shares (too unique) and more than 80% of the all users' shares (too popular). We use K-means++ clustering (K = 2 .. 10), which is an advanced version of K-means to improve selection of initial seeds. K-means++ effectively works for a very sparse matrix (user-link). To choose the optimal number of clusters, we use the "elbow" method.

Figure 3 shows that the marginal gain of the variance explained by the clusters dropped at that point. As a result, we have 8 clusters. We note that the K-means++ clustering method can be replaced with other clustering methods in various situations. Particularly, algorithms designed for streaming data (e.g. STREAM) might be better for hard real-time systems.

Fig. 3. Quality of clustering with different number of clusters.



What transforms a user cluster to a persona is to describe the representative demographic or detailed behavioral characteristics of a set of users of the cluster. We capture such information from the most discriminative links shared by the users in each cluster. We first list the top 100 domains from each cluster, discovering that there are large overlaps among clusters. We then remove from each cluster the domains that existed in another cluster in order to identify the relevant, unique, and impactful domains. This de-duplication results in the elimination of one cluster, leaving us with a set of 7 clusters. In other words, users in the eliminated cluster show similar behavior with users in other clusters, and thus it is not a meaningful cluster in order to generate a persona. As a result of de-duplication, each cluster can be characterized by domains shared only within that cluster. We call such extracted common interests "topics", as shown in Table 3.

TABLE II. MOST FREQUENTLY SHARED DOMAINS BY AJ+ USER USERS (DOMAINS WITH MORE THAN 30 THOUSAND SHARES).

Topics	No. (%)	Impactful Domains
Interested in family, and religion	25,825 (58.85%)	• www.beliefnet.com/ babyvote.com
Interested in entertainment	11,251 (25.63%)	• flickr.com • hulu.com
Interested in politics	4,058 (11.15%)	• huffingtonpost.com • nytimes.com
Interested in celebrities and casual reading	1,254 (2.86%)	• eonline.com • cosmopolitan.com
Interested in sciences and comics	542 (1.24%)	• comicbook.com • nerdist.com
Interested in odd assortment	515 (1.17%)	• tumblr.com • keek.com
Interested in quiz and games	436 (0.99%)	• quizdoo.com • memorado.com

Topic: "Interested in family and religion" (25,825 users, 58.85%): These users are serious news readers, with a broad range of interesting, but with leanings toward business, technology, and religion. Their viewpoints, based on their shared links, are mixed such as bible.com but also thespiritscience.net. Similarly, their political links are mixed, such as foxnews.com but also the huffingtonpost.com.

Topic: "Interested in entertainment" (11,251 users, 25.63%): These users are entertainment focused, especially music and sports, with interests in links such as stream.tv, grooveshark.com, flickr.com, nba.com, imdb.com, and hulu.com.

Topic: "Interested in politics" (4,058 users, 11.15%): The users in this group are news geeks and demonstrate an interest in political news. They also lean somewhat politically liberal, in the U.S. spectrum, sharing links like politico.com, bluenationreview.com, billmoyers.com, blackthen.com, occupydemocrats.com, and msnbc.com. These users also interact with domains concerning the MENA, including oamuslims.org, muslimvillage.com, and middleeastmonitor.com.

Topic: "Interested in celebrities and casual reading" (1,254 users, 2.86%): These users follow news about celebrities and mainly interested in general life articles, with a focus on an enriching personal life, sharing links from thrillist.com, pulptastic.com, bustle.com, smosh.com, refinery29.com, dose.com, and elephantjournal.com.

Topic: "Interested in science and comics" (542 users, 1.24%): The primary interests of the users in this group are science, animation, and comics.

Topic: "Interested in odd assortment" (515 users, 1.17%): These users interact with an assortment of links, generally somewhat edgy in topic, including tranquilmonkey.com, storify.com, and koreaboo.com. In addition, these users are somewhat international in viewpoint.

Topic: "Interested in games and quiz" (436 users, 0.99%): These users are interested in quizzes (e.g., brain training or teasing type) and games. Their political links relate to Jewish topics, including jerusalemonline.com and islamicsgnetworks.com.

In the above, we demonstrate that by using publicly available information on Facebook, we can extract user interests to generate clusters that feed into the generation of personas. Next, we show that we can generate richer description of personas by combining data from other social media platforms.

VI. ADDITIONAL DATA COLLECTION AND ANALYSIS (TWITTER)

Due to space constraints, we briefly discuss the use of Twitter data for automatic persona generation. Personas typically contain insights into the fictional character beyond behavioral data, including such attributes as occupation, hobbies, viewpoints, and life goals.

To integrate these persona aspects into our system, we leveraged the Twitter social media platform, specifically the Twitter bio, which is an aspect of Twitter profiles that users complete when registering on the platform. We leverage the Twitter bios to provide the non-behavioral aspects of the personas, specifically quotes constructed narrative including in the profiles.

We utilize two sets of Twitter users, (a) the general Twitter populations and (b) the population of Twitter users who follow AJ+'s Twitter account (@ajplus). For both sets of data, we identify Twitter users matching our interested demographics in terms of the set of matching criteria (e.g., user interests, gender, age, country, etc.). We then leveraged the matched Twitter users for construction of the appropriate quote, one from the general population and one from just AJ+ followers.

VII. ADDITIONAL DATA COLLECTION AND ANALYSIS (YOUTUBE)

Personas typically contain, along with behavioral characteristics, demographic information such as gender, age, and location, which can be captured from YouTube data. Due to space constraints, we briefly discuss the use of YouTube data for automatic persona generation. An example of an AJ+ YouTube video is shown in Figure 4, noting the likes/dislikes, shares, and views.

AJ+ YouTube channel analytics anonymously recorded user (e.g., gender, age, country location, and which site the user comes from) for each of AJ+ videos. We leverage this data to include demographic information into our automatic persona generation system. We correlated the demographic aspects with the topical classification of the video in question using a twenty item news topic taxonomy for the news videos. With the demographic data, we could locate similar Twitter accounts, and with the news topic, we could correlate with our Facebook clustering data.

Fig. 4. Example of YouTube Video from AJ+.

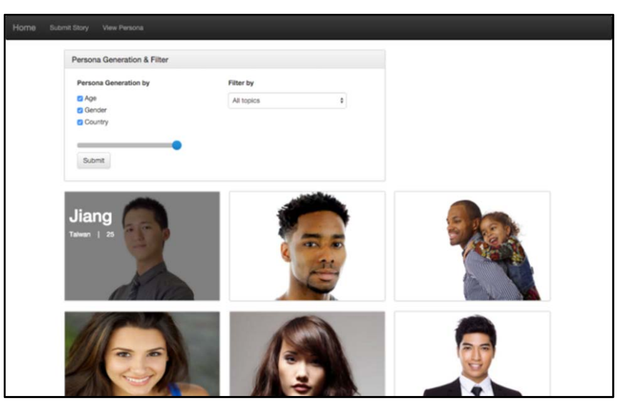
VIII. Automatic Persona Generation

We now present and discuss the current state of system development. Again, our research aim is to use actual user behavior, rapidly collected and analyzed, to generate personas in real-time. Therefore, we have developed the prototype for a system that automatically generates personas representing current readers of AJ+ news content based on actual user data from three social media channels (Facebook, YouTube, and Twitter) that updates these personas in real time based on changes in audience demographic, interests, or usage. We present screenshots of the automated personas generation system in Figures 5 and 6.

One can select the attributes that one desires to generate personas around, currently gender-age-country, and various levels of personas, which allows for varying levels of granularity. For a news organization such as AJ+, one can also filter the personas by topics, which for AJ+ is one of the twenty news article classifications.

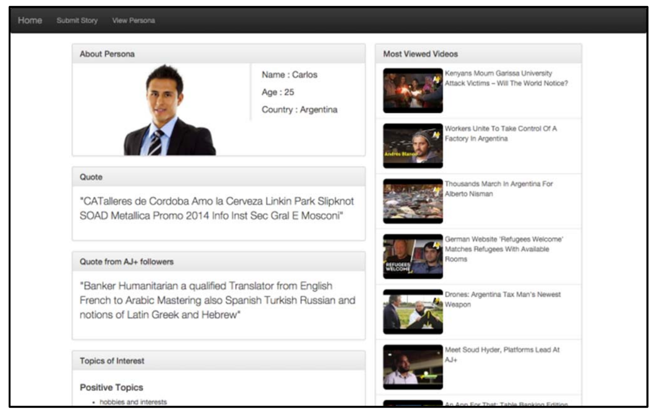
Selecting all attributes and the maximum number of personas generates the results shown in Figure 5.

Fig. 5. Screenshot of automated personas generation system screen for resulting personas.



As shown in Figure 5, multiple personas are presented, with names, ages, and pictures assigned in real time. The pictures, with all copyrights purchased, are gender, age, and ethnically appropriate. The names for the personas are also age appropriate (for example, if the fictitious person in the persona is a 55 year old Caucasian female, the name for the fictitious person is a common name for that year representing that gender and ethnicity). A mouse rollover causes the persona's name, age, and country to appear (see Figure 5, Jiang persona). Clicking on a persona causes the persona to appear, as shown in Figure 6.

Fig. 6. Screenshot of a single persona from the automated personas generation system.



As shown in Figure 6, the personas heading appears (e.g., Carlos). There are currently two quotes, one from the corresponding over Twitter population (e.g., Quote) and one from the specific AJ+ Twitter populations (e.g., Quote from AJ+ followers). Specifically, for AJ+'s news article focus, the top and bottom Topics of Interests are display. Finally, the screen displays hotlinks to the top ten most watched videos by this persona.

IX. Discussion and Future Direction

The current automatic persona generation system is quite robust, pulling in tens of thousands of social media records showing that one can use actual user behavior that is rapidly collected and analyzed in order to generate personas in real-time. However, there are several research and development fronts that we are pursuing.

A. Rich Personas Beyond User Interests

The biggest strength of our approach is that we benefit from actual user data, reducing time and cost for generating personas relative to traditional methods, and thus our approach is suitable for real-time persona generations. With our method, we do not need to carefully sample users for interviews to develop personas, but we can analyze and extract representative personas from millions of users by using publicly available online data.

Our research is a starting point for creating personas for a vast number of other applications and services without much manual efforts. Thus, if we can leverage more rich information of a user that is usually considered for defining personas, such as ethnicity, gender, hobbies, and socio-economic status, our methods and results would become more and more useful.

One possibility is to extract demographic information from shared links. For example, using the description of each of shared links, we can detect user languages. In our current dataset, for example, 55.2% (30,294) users share links in just one language, and 44.8% users share links in multiple languages. The most frequently used languages are English (31.98%), followed by German (5.69%), Spanish (5.02%), French (4.75%), Italian (3.46%), Indonesian (2.99%), Portuguese (2.94%), Dutch (2.94%), Tagalog (2.71%), and Afrikaans (2.69%).

Also, shared links could reveal user's economic status. Previous research has shown that affluent customers (high-end luxury product websites) and budget conscious customer (price aggregation or discount websites) can be distinguished by websites they visited [13]. Using these as examples, we are investigating extracting rich information beyond user interests from shared links.

B. Scalability of the System

One possible concern of our approach is the scalability of the system, as there could be hundreds of thousands to millions of user vectors. We have designed the architecture to scale up and handle this level of user data. For example, our system cumulatively accesses and stores the Facebook data using the Facebook API. The data collection process requires only a few number of HTTP requests for each user, and we only need 'newer data' once the initially collected. This means that the number of required requests does not increase dramatically and even decreases over time, and thus, we can also say our persona generation is near real-time.

C. Benefits for Journalists

Our persona currently provides initial demographic information and more rich behavioral details. Our approach is also derived from collaboration with journalists who actually benefit from the generated personas. They wish to have the realistic view of the actual users so that they can reach readers with better titles, content, and article framing. In this sense, our persona research offers a strong foundation for achieving this. Our resulting personas clearly describe what topics readers are interested in. They help journalists to search news topics that can potentially appeal to readers and to spin articles that attract those targeted readers.

X. CONCLUSION

We have taken the initial fruitful steps to move personas creation from a manual, time intensive process to that which is automated and in near real time. Currently, we integrate user data from three social media technologies, analyze this data, and generate personas beneficial to, in this case, journalists. We are continuing systems development to enhance features for personas selection and persona filtering. We also investigating leveraging other data sources, including other social media platforms and off-line sources [14] to provide richer demographic attributes, attitudinal characters, and other aspects for rounding out the generated personas.

REFERENCES

[1] A. Cooper, *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity*. US: Sam, 1999.

[2] T. Matthews, T. Judge, and S. Whittaker, "How Do Designers and User Experience Professionals Actually Perceive and Use Personas?," in *CHI2012*, 2012, pp. 1219–1228.

[3] N. Khalayli, S. Nyhus, K. Hamnes, and T. Terum, "Persona Based Rapid Usability Kick-Off," in *CHI 2007*, 2007, pp. 1771– 1776.

[4] J. McGinn and N. Kotamraju, "Data-driven Persona Development," in *CHI2008*, 2008, pp. 1521–1524.

[5] T. Judge, T. Matthews, and S. Whittaker, "Comparing collaboration and individual personas for the design and evaluation of collaboration software," in *CHI.2012*, 2012, pp. 1997-2000.

[6] J. E. Nieters, S. Ivaturi, and I. Ahmed, "Making Personas Memorable," in *CHI 2007*, 2007, pp. 1817–1824.

[7] B. J. Jansen, K. Sobel, and G. Cook, "Classifying Ecommerce Information Sharing Behaviour by Youths on Social Networking Sites," *Journal of Information Science,* vol. 37, pp. 120-136, 2011.

[8] S. Abbar, J. An, H. Kwak, Y. Messaoui, and J. Borge-Holthoefer, "Consumers and Suppliers: Attention asymmetries. A Case Study of Al Jazeera's News Coverage and Comments," in *Computational Journalism Symposium*, 2015.

[9] J. Roettgers. (2015, 30 July) How Al Jazeera's AJ+ Became One of the Biggest Video Publishers on Facebook. *Variety*. Available: http://goo.gl/76DZBP

[10] S. Faily and F. I., "Persona Cases: a Technique for Grounding Personas," in *CHI2011*, 2011, pp. 2267–2270.

[11] J. Sun, G. Wang, X. Cheng, and Y. Fu, Mining affective text to improve social media item recommendation," Information Processing & Management, vol. 51, pp. 444-457, 2015.

[12] F. Celli, A. Ghosh, F. Alam, and G. Riccardi, "In the mood for sharing contents: Emotions, personality and interaction styles in the diffusion of news," Information Processing & Management, vol. 52, pp. 93-98, 2016.

[13] A. Odlyzko, "Privacy, economics, and price discrimination onthe internet," in Electronic Commerce. 2003, pp. 355–366.

[14] P. Mukherjee, P. and B. J. Jansen, "Second screen interaction analysis for IRL events: Phase-category investigation of the super bowl 2015 social soundtrack," in ICICS2016, 2016, 13 - 19.