Title	Behavioural Analysis of Mobile Web Users
Authors(s)	Albatal, Rami; Briggs, Peter; Coyle, Maurice; Gavarini, Sebastian; Tomin, Johannes; Smyth, Barry
Publication date	2016-05-27
Publication information	Sprink, A., Riedel, G., Zhou, L., Teekens, L., Albatal, R. and Gurrin, C. (eds.). Proceedings of Measuring Behavior 2016: 10th International Conference on Methods and Techniques in Behavioral Research
Conference details	Measuring Behavior 2016: 10th International Conference on Methods and Techniques in Behavioral Research, Dublin, Ireland, 25-27 May 2016
Publisher	School of Computing, Dublin City University, The Insight Centre for Data Analytics, the University of Aberdeen and Noldus
Link to online version	http://www.measuringbehavior.org/mb2016/home
Item record/more information	http://hdl.handle.net/10197/8405

Behavioural Analysis of Mobile Web Users

R. Albatal*, P. Briggs, M. Coyle, S. Gavarini, J. Tomin and B. Smyth

HeyStaks Technologies Ltd., Ireland. rami.albatal@heystaks.com

Abstract

As smartphones become the predominant devices for accessing the web, understanding how individuals express their interests and interact with the web can have a great impact on several domains ranging from customer services to marketing and public policy. However, in order to better understand the web surfing behaviour and interests of mobile network subscribers, we need to look beyond the classic analytics that are based on location, internet usage and social networks. A more granular view of user behaviour and interests can be achieved by including more advanced analytics based on the content that the users are engaging with. In this paper we present a novel mobile web content analytics platform, HeyStaks, with the goal of filling the gap of granular content analysis for mobile user behavioural analytics.

Introduction

It is clear that smartphones are increasingly becoming our main content consumption devices when it comes to accessing the web. According to the latest forecast from Cisco [1]: "Mobile data traffic will grow three times faster than fixed traffic in the period 2014-2019, driven by more devices and users as well as faster networks. [...] The growth is driven by more mobile users, which are expected to grow from nearly 59 percent of the world population last year to 69% in 2019." Google has lately announced that "more Google searches take place on mobile devices than on computers in 10 countries."

With these facts in mind, the analysis of web content engagements on mobile devices is of great interest for many market and social players. For example, Mobile Network Operators (MNOs) are constantly seeking to optimise their services with better recommendations and customised plans for their users, advertisers are looking for more relevant and less aggressive targeting mechanisms, and governments and non-governmental organisations want to better understand people's interests, trends and social engagement at population level.

We argue that a far more granular and accurate level of user profiling can be achieved by adding the essential element of web content analysis. It is cumbersome to track users across websites using cookies, and doing so requires partnerships with many content publishers in order to achieve a reasonable level of coverage, But the service provider has a global view of the user's mobile web activity, and with the user's consent they can build a rich, anonymous profile of the user's personal preferences. This is where the HeyStaks platform comes in to bridge the gap and enable MNOs to build a richer user profile.

This paper outlines the HeyStaks platform, and some key results from an analysis of the Web usage patterns of mobile subscribers. The HeyStaks platform extracts and identifies behavioural patterns from this usage data to better enable marketers, advertisers, and other parties to create precise targeted campaigns for interested audiences.

Proceedings of Measuring Behavior 2016, (Dublin, Ireland, 25-27 May 2016).
Editors: A.J. Spink *et al.* www.measuringbehavior.org

The rest of this paper is organised as follows: we first start by presenting related state-of-the-art works, then we describe HeyStaks and it's methodology. Finally, we discuss some insights that we generated from a population of users from one of our partner MNOs.

Related Work

While the idea of investigating web logs is not new, there are few academic or industry research studies that combine the analysis of mobile web usage patterns with the automated profiling of mobile subscriber interests through the classification of the web content they access.

Many studies in the state-of-the-art have demonstrated the effectiveness of analysing web content and/or the URLs of the websites that users engage with. Hofman & Sirer[2] analysed the Web histories (by categorising the web pages into 5 categories) of 250,000 anonymized individuals with user-level demographic information. They examined how the surfing behavior changes as individual spend more time online, how it depends on educational background, and how browsing histories can be used to infer user's attributes. In [3] a method for predicting the dwell time on Web pages was proposed based on features related to the content (words), the HTML tags and measurements on the page size, heights and width and others. In [4], it was possible to predict some of the demographic information of users using their URLs visits patterns. Authors in [5] applied a clustering technique on URLs visited by the users to capture the common interests of different types of users.

When looking at mobile user behavioural analytics, most of the research has focused on one or a combination of three dimensions: social networks, mobility, and Internet usage metrics. For example, the study in [6] focused on the analytics of the user behaviour change over different networks using network traffic data. In [7], sub-communities of similar users are identified based on mobility and network traffic pattern analysis. Anindya & Sang [8] measure User content generation and usage behaviors based on factors related to calling patterns and mobility. A data mining method (SMAP-Mine) was proposed in [9] to discover the sequential movement patterns associated with requested services (e.g. restaurants, theatre).

Few studies used the domain names of the websites visited by a mobile user to perform behavioural and interest analysis. The authors in [10], proposes a probabilistic model hat combines the user location and the user interest profile generated by applying Latent Dirichlet Allocation over a bag-of-websites (domain names) representation, the model was used to perform collective behavioural analysis for mobile usage prediction and service recommendation. In [11], a profiling and recommendation approach based on fuzzy clustering is applied on the URLs visited.

The studies, and many others, can be very beneficial in enhancing the quality of user targeting and recommendation systems, however none of them apply advanced analytics on the content of the requested web pages. This is where the HeyStaks platform contributes to this domain by carrying out a detailed content analysis on the visited web pages to provide a richer user profile.

In the next section we will provide a high-level overview of HeyStaks and its content analysis methodology.

HeyStaks Big Data Platform for Mobile Web Behavioural Analytics

Figure 1. shows a high-level overview of the HeyStaks platform for mobile Web behavioural analytics.

The platform constitutes of two parts:

- First part (left side of Figure 1.): installed in the MNO Network. This part is mainly responsible for building, managing and updating the subscribers profiles. URL logs containing timestamped URLs with user ID and Location are read and parsed by the *Update API*. The Profile Manager module takes charge of aggregating the information related to each subscriber, and it requests the topics related to each URL from the *Topic Manager* module. The user profiles are then saved in a database that can be accessed via the *Access API*. The profiles can be further analysed by the *Community Manager* module.
- Second part (right side of Figure 1.): this part is outside the MNO network and contains the core analysis capabilities of the HeyStaks platform. It receives requests from the *Topic Manager* module that ask for the topics related to new URLs that the system has not encountered before. In this case, the communication is conducted via a secure connection. The cloud service will fetch the content of the URL and analyse it via Machine Learning models. The Machine Learning models map the content of a URL into one or several topics in a taxonomy.

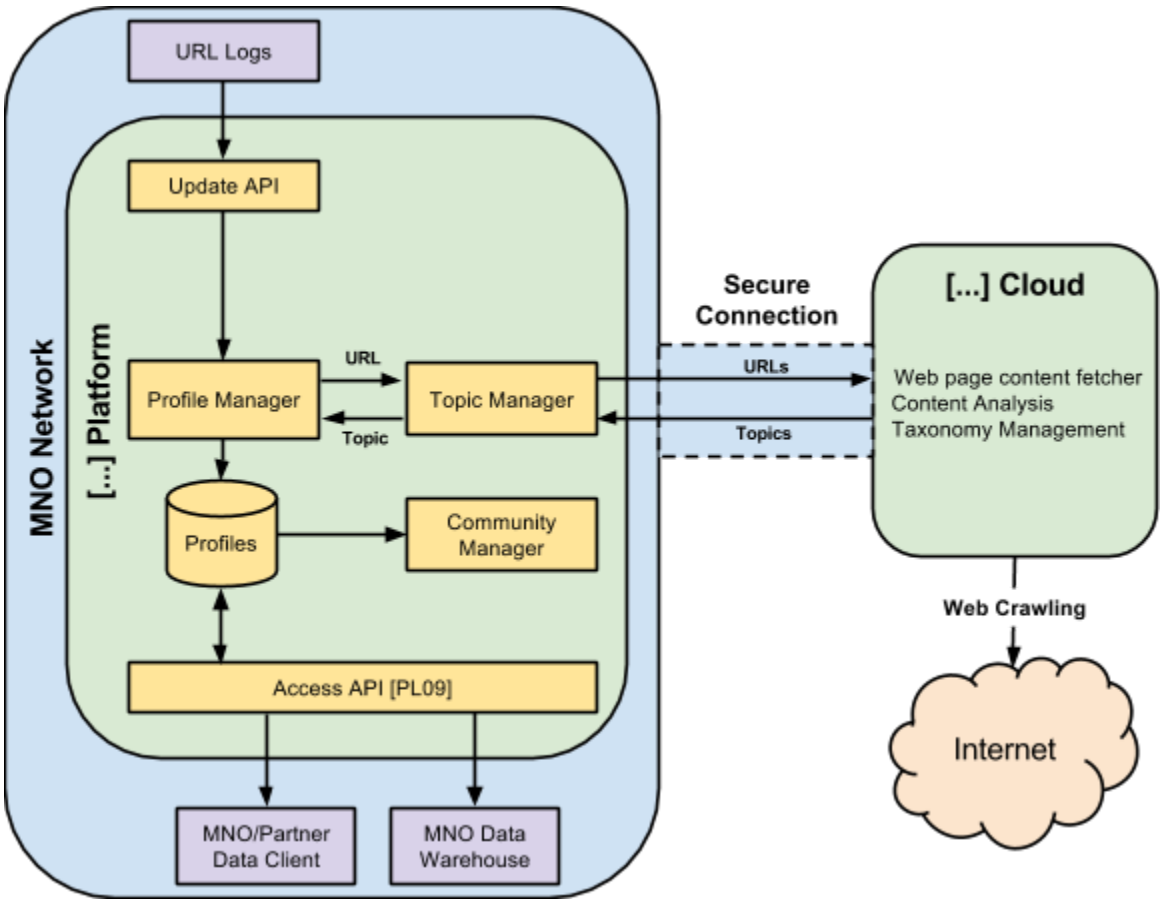


Figure 1. Overview of the HeyStaks platform for mobile Web behavioural analytics

Next we will provide some details about the modules that affect or contribute to the behavioural modelling of users.

HeyStaks Taxonomy Management

HeyStaks maintains a Taxonomy of categories that are used to classify the high-level interests of end users, and which is based on the IAB Quality Assurance Guidelines Taxonomy [12].

While the HeyStaks Taxonomy has been carefully designed to meet the needs of most use-cases for HeyStaks Profile data, some customers have specific requirements in this area. To accommodate these differing use-cases, HeyStaks supports alterations to the standard Taxonomy on a per-customer basis to cater for more localised interest topics. The Taxonomy may also be extended to allow for alternative use-cases for HeyStaks Profiles. For example, an eCommerce-focussed taxonomy could be constructed to meet the profiling requirements of an online store.

HeyStaks Web Content Analysis

The techniques used to determine the topics of the web pages include a structural analysis of the page content, link graph analysis of the pages, text relevance analysis, page term statistics, and intent analysis of search queries and page characteristics. The variety of content analysis techniques used ensure a high quality mapping between the web pages visited by the users and the topics in the HeyStaks taxonomy, and allow the system to perform well in a multilingual web environment.

First, the text processing step applies a standard stop-words list (available on http://www.ranks.nl/stopwords). Next, page features are extracted and mapped to an index using a hash function. This allows term frequencies to be calculated for the page features based on the mapped indices. This approach avoids the need to compute a global term-to-index map, which can be expensive for a large corpus. Along with Term Frequency/Inverted Document Frequency (TF-IDF) features [14], Machine Learning algorithms are used to analyse the textual content of the web pages. Specifically, Latent Dirichlet Allocation (LDA) [13] is employed for feature extraction. Textual features are then used as input for Naive Bayes and Logistic Regression multi-class classifiers. To train these classification algorithms, a ground-truth dataset is automatically constructed using common web search engines and open knowledge-bases such as WikiPedia. Relevant pages for each topic in the taxonomy are collected, then the pages are processed and features are extracted and used as input for the training algorithm. The generated classification models are then used to estimate which taxonomy topics are most likely to be related to a web page, then the pages are filtered by comparing them to reference pages on WikiPedia. The classification quality is measured using an "Accuracy@1" metric, and it is calculated over a standard 80%-20% split of the data into training and validation sets respectively. The system achieves an Accuracy@1 of 78% for pages with English content, and classifiers for several additional languages are currently being evaluated.

HeyStaks Community Manager

After analysing the web pages that a user visit, the results of the analysis are aggregated in the user profile and saved in *Profiles* database. The *Community Manager* module carries out offline processing on the HeyStaks *Profiles* database to identify communities of similar profiles within the MNO's subscriber population. Sets of similar users are computed using several clustering techniques, and these sets are then used to generate community-level profiles that aggregate the information in the underlying profiles, and also as input to the processing carried out by the Interest Inference Manager. In the next section we will present some of the insights that are produced by the HeyStaks *Community Manager* module.

Insights

Let's take an example of one day of web interaction data that belongs to a sample of 45K unique users from a mobile operator with approximately 3M subscribers. We will present 2 types of insights that we can generate from this sample.

A. Community Identification

In Table 1. below we see 3 examples of communities that were identified within the Web usage logs of a mobile operator. There are clear patterns of overlap in each community, indicating that the members are quite similar.

	Community 1	Community 2	Community 3
Primary interests	Business News	Clothing	Tourism
	Marketing	Women's Fashion	Accommodation
	Business Services	Footwear	Resorts
	Business Operations	Beauty	Vacation
Secondary interests	Business Logistics	Fashion Accessories	Social Science
	Wholesale	Health & Fitness Products	Computer programming
	Education	Men's Fashion	Travel Agencies
	Business Training	Cosmetics	Sport Events
	Business Associations	Fashion Designers & Luxury	Multimedia
	Newspapers & Mags	Skin Care	Computer Science
	Commerce	Shopping collectibles	Video Games
	Movies	Tourism	Consumer Electronics
	Tourism	Hair Styling	Religion

Table 1. Mobile Users Communities detected based on content analysis of Web usage

This fully automated processing of the user profiles reveals interesting communities that can be understood, described and can even be used to infer demographic information about the subscribers. It is clear from looking at the short profile excerpts in Table 1. that certain trends emerge - for example when we look at the 'Primary interests' of the communities, it looks like '**Community 1**' consists of people who are interested in Business, '**Community 2**' contains people who are interested in fashion and beauty, and '**Community 3**' contains people who are interested in travelling.

B. Web Usage Behaviour

Using the *Community Manager* module, we can explore the web usage behaviour of the totality of the user population or we can decompose it into the usage behaviour of any discovered community. The graph in Figure 2. presents the global web usage behaviour in terms of number of engagements distributed over the 24 hours of a typical day.

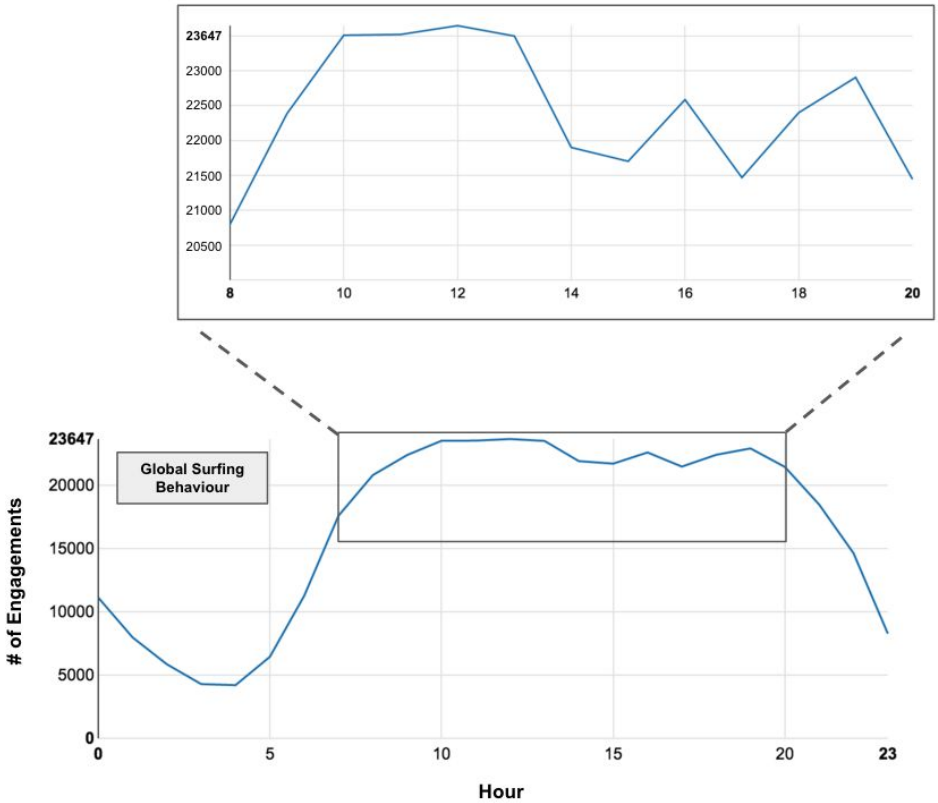
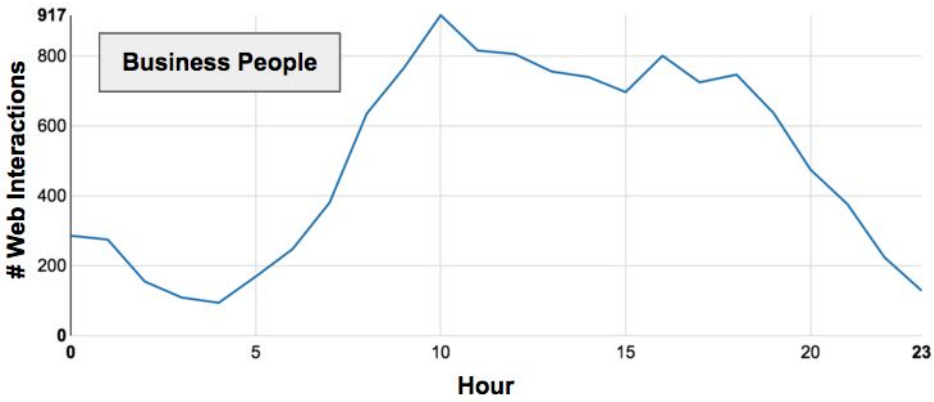


Figure 2. Global Web Usage Behaviour of a Population of MNO subscriber

What we notice is that across the entire user population there is a consistent level of usage during peak hours (between 8am and 8pm), and that there are a couple of particularly popular hours (12pm, 4pm and 7pm) that appear as peaks compared to their adjacent hours.

When looking at the communities in the Web Usage Behaviour Graphs in Figure 3, we can easily notice the differences. This information can play a crucial role in identifying the optimal time of day to start an advertising campaign, for example. The advertiser can now harmonise the timing of his targeted ads with the interest graphs of his target audience so that people see relevant ads when they are most likely to be useful and least likely to be an annoyance.



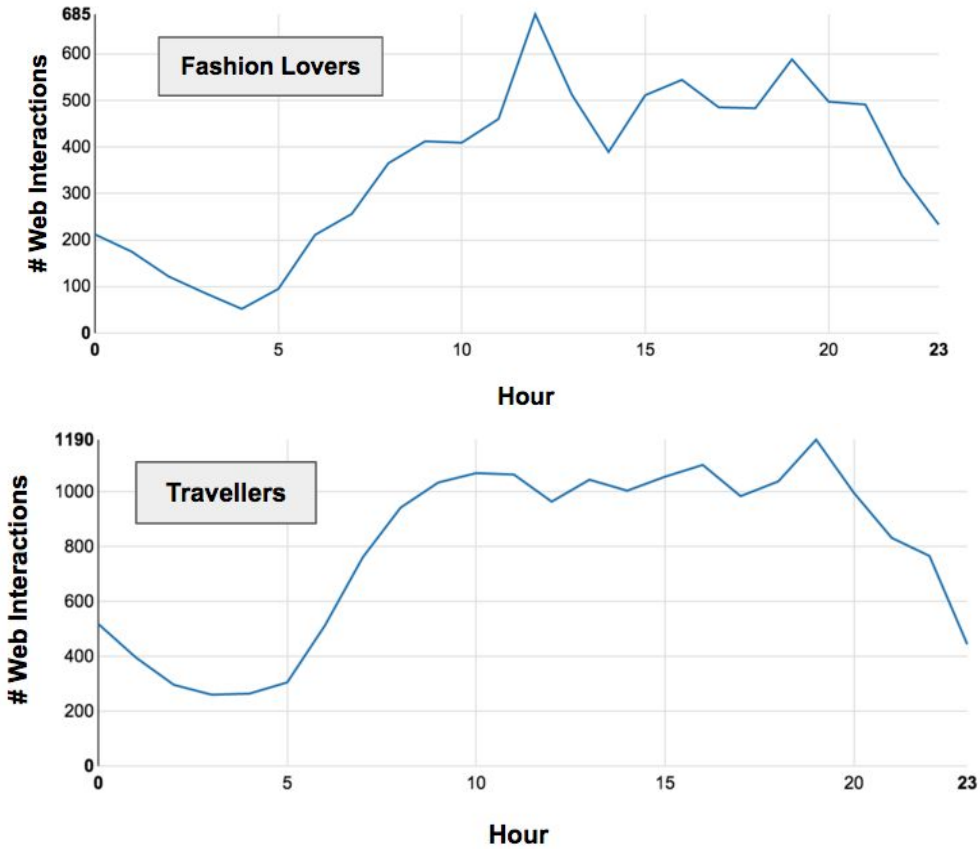


Figure 3. Surfing Behaviour Graphs of different MNO subscribers communities

It is interesting to see how different communities behave in different ways throughout the day. 'Business People' tend to read the business news while having their morning coffee or when they arrive at work. Many of our 'Fashion Lovers' check the Fashion websites and content during lunch, hoping to have some time in the evening to do some shopping. The 'Fashion Lovers' is a group that exhibit an intent to buy clothing, so they may respond positively to ads or offers for clothing deals around lunch time. After getting home from work, the 'Travellers' like to plan their trips and check promising holiday destinations in the evening, so this is the right time to target them with holiday ads and tickets promotions.

Ethical aspect of the study

The web usage data that was analysed in the course of this study was pseudonymized, which involved pre-processing the data to encrypt the user identifiers with a hashing function. The study was conducted by calculating aggregate statistics over the usage patterns and interests of the entire user population, and did not involve the analysis of individual users' data.

Conclusion

In this paper we presented some modules of the HeyStaks Mobile Data Analytics Platform. We have described the main modules that are used to analyse the web pages, to construct user profiles, and to analyse the collective behaviour of users communities. The interesting behavioural insights generated by the HeyStaks platform can allow MNOs and other interested parties to know **who** their subscribers are, **how** they engage with the web and **when** they are likely to be interested in specific topics.

References

1. “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019 White Paper”. Cisco Systems. February 3, 2015. Link. Published 3 February 2015. Accessed 27 January 2016.

2. Goel, S.; Hofman, J. M. & Sirer, M. I. (2012). Who Does What on the Web: A Large-Scale Study of Browsing Behavior. *In proceedings of the 6th International AAAI Conference on Weblogs and Social Media* (Dublin, June 2012), 1-8.

3. Liu, C., White, R.W., Dumais, S. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* - SIGIR '10 (New York, NY, USA, 2010), 379-386.

4. Adar, E., Adamic, L. A., Chen, F. R. (2007). User profile classification by web usage analysis. Google Patents. http://www.google.com/patents/US7162522.

5. Mobasher, B., Cooley, R., Srivastava, J. (1999). Creating adaptive web sites through usage-based clustering of urls. In Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange - KDEX 1999 (Chicago, Illinois, USA, 1999), 19-25.

6. Tsompanidis, I., Zahran, A.H., Sreenan C. (2014). Mobile Network Traffic: a User Behaviour Model. *In Proceedings of 7th IFIP Wireless and Mobile Networking Conference* (Portugal, May 2014), 1-8.

7. Tang, D., Baker, M. (2000). Analysis of a local-area wireless network. In *Proceedings of the 6th annual international conference on Mobile computing and networking* (NY, USA, 2000), 1-10.

8. Anindya, G., Sang P.H. (2011). An Empirical Analysis of User Content Generation and Usage Behavior on the Mobile Internet. *Management Science*, 57(9), 2011, 1671-1691.

9. Tseng, V.S., Lin, K.W. (2006). Efficient mining and prediction of user behavior patterns in mobile web systems. Inf. Softw. Technol. 48, 6 (June 2006), 357-369.

10. Yua, B., Xu, B., Wu, C., Ma, Y. (2014). Mobile Web User Behavior Modeling. In Proceedings of the 15th International Conference on Web Information Systems Engineering - Wise 2014 (Thessaloniki, Greece, 2014), 388-397.

11. Phatak, D., Mulvaney, R. (2002). Clustering for personalized mobile web usage. In Proceedings of the 2002 IEEE International Conference on Fuzzy Systems - FUZZ-IEEE 2002 (Budapest, Hungary, 2002), 705-710.

12. IAB Quality Assurance Guidelines (QAG) Taxonomy. http://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag-taxonomy/. Accessed 27 January 2016.

13. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993-1022.

14. tf-idf. https://en.wikipedia.org/wiki/Tf-idf. Accessed 27 January 2016.