A Survey in The Application of Data Mining in Web Page Personalization & Consideration for Privacy in Digital Personalization

Israel Nolazco
Introduction to Data Mining and Analytics
Lewis University
Romeoville, Illinois
israelnolazco@lewisu.edu

Paul Richardson
Introduction to Data Mining and Analytics
Lewis University
Romeoville, Illinois
paulmrichardson@lewisu.edu

***Abstract*—Digital personalization for web pages is a process in which direct and indirect information is gathered from the consumer. The goal is to create a unique experience tailored to the user to further expand the relevancy of the webpage. However, given the lack thereof technical expertise from the consumer side. This has created a divide between companies collecting data and consumers afraid of the misuse of their information. Specifically with the rise of electronic commerce (e-commerce) companies have access to a potentially limitless pool of data. This document will explore the different sources and techniques web pages take before collecting data. Additionally, This document will also present the process and techniques Data Analyst go trough before presenting their conclusions and lastly, this document will present ethical concerns with each step and postulate a potential solution to such conundrum.**

***Index Terms*—Big Data Analytics, E-commerce, Web Personalization, Ethics**

I. Introduction

The rise of electronic commerce (e-commerce) has added a new dimension to how businesses communicate with their consumers. In the past, businesses conducted market research, surveys, and business plans to organize, strategize, and execute on new ventures and campaigns. This process can be slow, expensive, and potentially unrewarding. E-commerce, however, allows a business to reach millions of consumers and potential customers around the world. The proliferation of e-commerce and digital solutions comes the creation and capture of consumer data that is unprecedented and ushering in a new set of social complexities balancing how to leverage data in support of capitalism and the extent to which individual user privacy protection should be applied. To exploit this data in pursuit of capitalism, 91% of fortune 1000 companies are investing in analytics projects [1]. Specifically, companies are now putting greater emphasis on the personalization of the digital consumer experience as a differentiator that helps create a competitive advantage in their markets [2].

This document intends to provide an overview and illustrative examples of how data is driving the ability to personalization the user experience in e-commerce with the use of data mining techniques. We refer to this personalization of the user experience in the application of e-commerce as digital personalization. Specifically, in this document, we define digital personalization. We look at the application of digital personalization across digital domains such as web pages, web content, and digital marketing. Additionally, we further discuss some of the ethical concerns for personalization as it relates to user privacy.

Identify applicable funding agency here. If none, delete this.

II. Definition of Digital Personalization

Some of the prevailing definitions of personalization are as follows. *Personalization is the ability to provide content and services tailored to individuals based on knowledge about their preferences and behavior* [Paul Hagen, Forrester Research, 1999].

Personalization is the use of technology and customer information to tailor electronic commerce interactions between a business and each individual customer. Using information either previously obtained or provided in real time about the customer, the exchange between the parties is altered to fit that customers stated needs, as well as needs perceived by the business based on the available customer information [4]".

Personalization is the capability to customize customer communication based on knowledge preferences and behaviors at the time of interaction [with the customer] [Jill Dyche, Baseline Consulting, 2002]; and By leveraging customer reactions to personalized products and services, companies continuously improve their personalization processes through an iterative feedback loop resulting in the virtuous cycle of personalization [4]".

Personalization is about building customer loyalty by building a meaningful one-to-one relationship; by understanding the needs of each individual and helping satisfy a goal that efficiently and knowledgeably addresses each individuals need in a given context [4]".

All of the definitions stated above are adequate for our analysis. However, they speak in a general sense to a practice that has been around since the advent of capitalism. For that reason, we distinguish between general consumer personalization and

978-1-7281-4387-3/19/$31.00 © 2019 IEEE

digital personalization. *Digital personalization is the ability or attempts to customize an online user experience to meet individual customer preferences through the personalization of web pages*. We will also discuss the social implications of accessing collecting and using user information to support personalization.

Another distinction worth considering is the difference between digital personalization and its relationship with customer segmentation. Customer segmentation may use data clustering techniques to help identify broader segments of customer groups. Digital personalization is the ability to specifically provide individual user experiences based on their own set of preferences. Digital personalization is attempting to create a one to one relationship while segmentation only accomplishes a one to many relationships with customers or potential customers.

Segmentation is the process of bucketing or separating prospects into similar aggregate groupings. Personalization is the ultimate goal of customizing the user's experience specific to their individual preferences based on information and insights about individual users. Personalization represents a level deeper beyond grouping to the particular prospect.

A. The Digital Personalization Process

Digital personalization is an iterative process defined by three stages of understanding, delivery, and measurement [4]. These three stages exist in every niche of digital personalization. In the scope of digital customization exists the application of content-based filtering systems, collaborative filtering systems, and rule-based filtering systems. These approaches and methods create the foundations of recommendation systems for products and page content. Web-based personalization focus is on the mining of user information that can aid in creating unique web page experiences focused on look and feel, navigation, and structure [8].

The technical implementation of this process is beyond the scope of this paper. The critical understanding is that for implementation of digital personalization methods for data capture must exist within the digital solution or digital ecosystem for providing digital consumers a personalized experience. That data is then analyzed to provide and deliver customized experiences. The process is refined with the continuous measure of effectiveness through user satisfaction scores, revenue, or a combination of measures appropriate to the desired outcomes [4] [5].

The technical implementation of web page personalization is currently the significant opportunity and challenge open for exploitation. Whether a new algorithm or technological advancement or the combination of both appear in the future, it is clear this area is open for innovation.

III. Digital Personalization in Web Pages

Web personalization implies the delivery of dynamic and personalized content, such as textual elements, links, advertisements, product recommendations, etc., that are customized to needs or interests of a particular user or a segment of users. The process of personalization involves data collection and preprocessing phase in which the information pertaining to user interests is obtained and preprocessed and a discovery phase in which user profiles are constructed from the data collected [6].

The personalization of web pages allows for and creates the potential to customize through the highlighting of existing hyperlinks, the dynamic insertion of new hyperlinks that seem to be of interest for the current user, or even the creation of new index pages [8]. Web personalization is and can also be extended to the size of the text, the layout of the page content, the color of fonts all based on as users individual preferences.

Web personalization is a fascinating new development in digital technologies still being maturing as the challenge to fully capitalize on the concept is still wide open to innovation. Nevertheless, this issue is not usually exploited so as to meet specific users needs, making the Web pages personalization (both in terms of content and shape) an interesting and unsolved challenge [7].

The overall process of web personalization consists of five modules, namely: user profiling, log analysis and web usage mining, information acquisition, content management, and web site publishing [6] [8]. Figure 1 and figure 2 provide a couple of general abstractions of the personalization process.

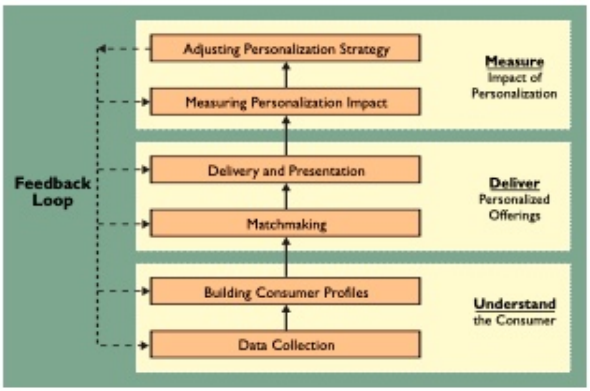

Fig. 1. Generalized Personalization Process [4]

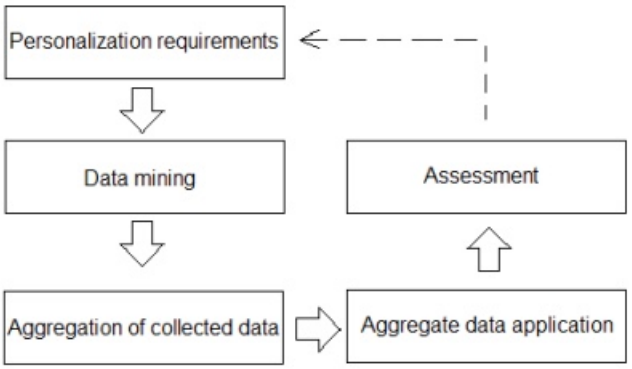

Fig. 2. Generalized Personalization Process [6]

The research and limited implementation consists of using web server logs, IP address, cookies, and authentication details to identify users, their usage patterns, and then publishing websites with structure and content based on that analysis. Popular predictive analytics methods such as classification, clustering, rule association, sequential pattern discovery, and are other statistical methods are used to tailor the user experience according to the results of the analysis. The challenge in web personalization is primarily in the capturing and preprocessing of the data. Server logs are a rich source of data but not designed for the application of web page personalization solutions [4] [5] [9].

Some of the research and developed prototypes have aimed at augmenting their user information through the development of interactive browser extensions or other client-side solutions. These extensions provide direct feedback and preference information from the user that allows for the website to make direct changes to the page text size, zoom level, font type, text spacing, background and foreground colors, and language translation [7] [10].

Other research has suggested specialized algorithms at every step of the process and an overall methodology to bring web personalization to fruition [11]. This approach like many of the papers also discusses the challenges with the pre-processing of server logs. It reinforces the concept that server logs contain a plethora of data, but that their is a real challenge in pulling out the pertinent information. The challenges exits in the volume and format. The actual records in the log files that provide insight and those that are just noise is also up for determination.

Another suggested solution that the authors of this paper theorized is the adoption of common frameworks to logs, HTML tags, and XML tags and other critical structures in the solution stack that can enable the capture or the curation of the metadata necessary. Identifying a framework would create an avenue for a solution that moves beyond trying to pull the essential meta-data from a digital ecosystem not designed with personalization and data mining in mind.

Current implementations of web personalization vary in form and are rudimentary. The potential is immense for a high level of web personalization. In the meantime, private companies like Hubspot are using propriety technology to meet this growing need and building an effective consumer strategy. Hubspot is currently able to deliver personalized user content, personalized web pages, and customized web page elements with embedded HTML tags and integration into their service offerings [12]. Companies like Hubspot have created platforms with the intent of robust digital personalization for use across the customer journey.

For clarification, content personalization is a subset of digital personalization and arguably, is not specifically web personalization. There is a relationship between the concept of webpage personalization and the creation of personalized user experience and related content for the user. The concept of recommendation systems as it relates to digital personalization and content personalization is the subcategory for this relationship [13].

IV. Ethical Concerns for Digital Personalization

There are two potential feelings when it comes to technology. There are those who wish a system was already set up and there are those who wish to customize every single detail of their experience. It is also within these two groups that we see similar feelings when it comes to data personalization. The group that would like to control every detail of their experience could potentially be identified as Technophiles. Users who push technological advances to their limits. The group that typically wishes to have a predefined set up are those who lack expertise or interest in adopting a new piece of technology. This group could be identified as a nonpower user. With this insight, it is with no wonder companies spend tremendous amounts of time to ensure their customization tools truly becomes personal. Power users could make the most of their customization options while nonpower users still use their product because of its quality. It is then why companies at any level have invested heavily in data collection. In order to make bricks one must gather clay and with that sentiment companies have tried to gather data as quickly and efficiently as possible thus sacrificing some consumer privacy or purchasing consumer information. To an emerging or an established company that focuses on profits; both situations are liable. As clearly stated by Bhargav Mantha ...And enterprises should continually monitor the success of their data usage and implementation to ensure theyre getting what they need out of it. Given this sentiment is it without a doubt consumer have become wary about taking advantage of the next big thing. In the past power users typically jump with the next trend, but if they are however informed about the misuse of data from a company. Then, the roles switches; powers users become defensive and unwilling to yield their data while the nonpower user may continue to use the service due to the lack thereof information about the companys behavior and use of personal data [8] [14].

Certainly, companies are aware of the potential harm it may suffer from misusing or improperly collect/handle data from its consumers. However, there is more than meet the eyes. In this document, we are going to identify the individual/team that analysis and handles consumer data by the term Data Analyst. In order to properly make use of the data collected a Data Analyst is typically call in order to interpret/make use of the data. After all, Data Analyst is there to give meaning to data and interpret to the best of their abilities the possible meaning behind such a piece of information. The worst-case scenario a Data Analyst faces is called the cold-start problem [16]. This term typically refers to the lack of previous user data. This is typically common for emerging fields of study or new lines of business. This is potentially one of the reasons companies are forced to purchase consumer data. Now, let us consider a situation where a Data Analyst would not face such a predicament and instead is handed a set of data from the company he/she works for, that is to say, web data analytics is been ongoing and now the analyst is tasked to make use of

it. The analyst in question could potentially use data mining techniques such as clustering, Bayesian classifiers, decision trees, nearest neighbor, ect.. To create a personalized web experience for a group. This might be satisfactory to the nonpower user since its concern is only to what is affecting him/her in a broader sense. As for the power user, she/he might feel hungry for more information and personalize content. A recommendation algorithm could then take in place to predict consumer behavior and show content that may appeal directly to the user, but at that point, we are now facing another ethical dilemma. By not showing certain bits of data that might be within the topic of the content; we are then excluding and inadvertently withholding information. This is also known as the Filter Bubble [15]; see figure 3.

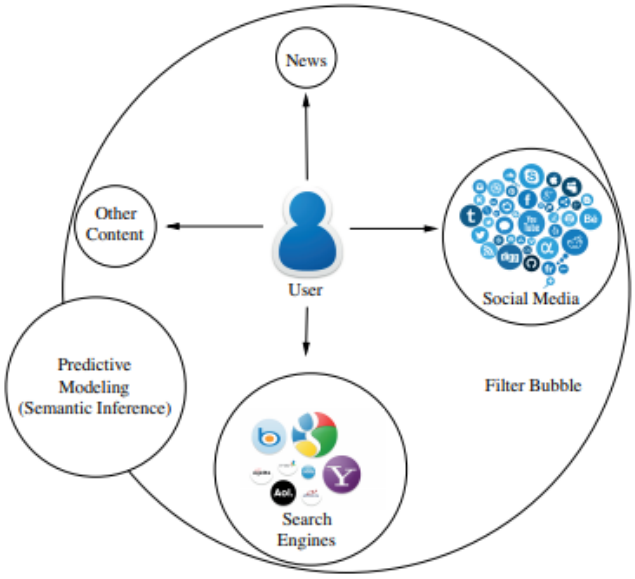

Fig. 3. The image represents the Filter Bubble and possible methods to break such a bubble

Interesting enough the Turkish Journal of Electrical Engineering and Computer Sciences provided a small guideline to break web users out of their filter bubble and in essence create a broader consumer experience. The recommended guideline included feeding the users profile with non-personalized content or recommend unrelated content. However, this is presented more of an option rather than a mandate. [15].

Let us consider our initial two groups, the power, and nonpower users. The power user will have a well-tailored web experience with data that is only relevant to him/her. The nonpower will also have that same experience with the caveat that web personalization will have fewer data to work. We will see both filter bubbles with both users, but one will potentially be more pleased with the content than the other. The power user as mentioned before is willing to provide as much data about itself to get that tailored experience. The nonpower user may not be so forward and its filter bubble by turn will reflect this lack of data. It is crucial to understand to a nonpower user the web is nothing more than just a tool and just like a hammer does not need to a name before it gets smashing; the sentiment is the same.

This document thus far showed the techniques Data Analyst uses and potential ways web personalization could affect users if they go unchecked. Now, let us consider another perspective. How the data is collected and the due diligence necessary to collect said data. The lack of accountability and understanding of the inherent risk of the misuse of user data have raised a few new questions. How much data is truly enough for an analyst? And what point of the analysis will we cross consumer privacy? Let us consider a case outside of Web Personalization in order to see the potential risk of mishandling data. The International Rescue Committee (IRC) in recent years has seen raising safety and confidentiality concerns. Influential donors are making increased demands for individual survivors' information and have misplaced confidence in how they might be able to use that information. Often exacerbated by donors harassing service providers, including by threatening to withdraw funding if the data are withheld. This, in turn, has put those in need of the organization's help in a vulnerable place. This is a clear case of an organization crossing consumer privacy. There have been calls to implement frameworks similar to that of the General Data Protection Regulation (GDPR), a legal framework that sets guidelines for the collection and processing of personal information from individuals who live in the European Union (EU). The framework could potentially aid Web Personalization analytics in ensuring the safety and dignity of clients as the first priority, including by extensively regulating data-sharing protocols to ensure confidentiality, consent, and related protections. Restore and defend the definitions of consent and confidentiality, recognizing that having a 'mandate' does not replace consent and cannot be used as a specific reason for sharing data and lastly create an internal body to identify and hold accountable for any misusage of said data [16] [17].

Although, GDPR seems to have a solid framework regarding data privacy. Web Personalization could still face human bias as a result of Data Analyst personal beliefs. Take into consideration the presence of online extremist on Twitter. These radical groups utilize Web Personalization to spread their message to those who are truly interested in the message. In essence they take advantage of the portal by recruiting, promoting, and increase participation across their followers While it is rational to let law enforcement officials police and investigate potential links to radical groups in social media. There have been cases where law enforcement agents target social groups based on lose connections. Ethical groups in the past have been prosecuted based on apparent connection rather than the actual crime and this creates an environment where personal freedoms could potentially be violated. A person has the right to express their personal thoughts, but in essence, it should not create chaos from its message. In this situation, Data Analysts that create the Web Personalization experience for its consumers should also be part of police investigations by providing control groups and random data to test their

theories. In essence, law enforcement should adopt scientific research techniques in order to keep personal freedoms and ensure appropriate investigation in their cases. Rather than considering data points as just data, consider them human subjects. Much like in the medical field a certain level of privacy is kept until it comes in the way for the health of the patient. That is to say, lets keep consumer behavior and expression private until it becomes a potential threat to the community by doing ethical research. Data Analyst should not see themselves as anything more than calculating machines, but rather gatekeepers of human rights and integrity [18].

V. Conclusion

"A feedback loop incrementally produces higher personalization. We all love personalization and the possibilities thanks to ML are limitless. However, managers need to be ready to address the question, How much is too much?" Andrew McStay, Professor at Bangor University [19].

The use of digital personalization for web pages has some immense potential, with some current challenges actively being addressed. In this paper, we discussed web page personalization and provided some considerations and perspectives that must be accounted for as it relates to user privacy. The topic of digital personalization is deeply rooted in online marketing and e-commerce. There are several other areas in the digital marketing domain that personalization is dramatically influencing and changing. According to Scott Brinker, VP Platform Ecosystem, HubSpot, "AI algorithms [and machine learning] are going to be deeply embedded at every layer of what the marketing system is [18]." The systems include web pages, content generation, keyword research, email, and product recommendations. Additionally, in support of the marketing channels is the concepts of lead generation and customer re-engagement. All of these marketing domains and channels have the potential, if not already, to deploy a level of digital personalization by deploying data mining.

References

[1] T. Le and S.-Y. Liaw, Effects of Pros and Cons of Applying Big Data Analytics to Consumers Responses in an E-Commerce Context, Sustainability, vol. 9, no. 5, p. 798, May 2017 [Online]. Available: [Accessed: 1-Sept-2019]

[2] M. Kupec, Web Personalization as a Corporate Digital Agenda Process, Marketing Identity, 01-Jan-1970. [Online]. Available: https://www.ceeol.com/search/article-detail?id=773284. [Accessed: 13-Oct-2019].

[3] C. Goward, Build the most effective personalization strategy, WiderFunnel Conversion Optimization, 01-Nov-2018. [Online]. Available: https://www.widerfunnel.com/personalization-strategy/. [Accessed: 13-Oct-2019].

[4] G. Adomavicius and A. Tuzhilin, Personalization Technologies: A Process-Oriented Perspective. [Online]. Available: http://people.ischool.berkeley.edu/ glushko/IS243Readings/Adomavicius.pdf.

[5] M. Kupec, WEB PERSONALIZATION AS A CORPORATE DIGITAL AGENDA PROCESS, FMK. [Online]. Available: https://fmk.sk/download/Marketing-Identity-Digital-Mirrors-I.pdf.

[6] S. Sharma and V. Rana, Web Personalization through Semantic Annotation System , Research India Publications. [Online]. Available: https://www.ripublication.com/acst17/acstv10n6_14.pdf.

[7] Mirri, Silvia, Catia Prandi and Paola Salomoni. Experiential Adaptation to Provide User-Centered Web Content Personalization. (2013).

[8] Magdalini Eirinaki and M. Vazirgiannis. "Web Mining for Web Personalization" ACM Transactions on Internet Technology (2003): 1-27. doi:10.1145/643477.643478

[9] M. Baglioni, U. Ferrara, A. A. Romei, S. Ruggieri, and F. Turini, Preprocessing and Mining Web Log Data for Web Personalization, University of Pisa. [Online]. Available: http://pages.di.unipi.it/ruggieri/Papers/aiia2003.pdf.

[10] H. Lu, Q. Luo, and Y. K. Shun, Extending a Web Browser with Client-Side Mining, SpringerLink, 23-Apr-2003. [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-36901-5_18. [Accessed: 13-Oct-2019].

[11] Rathi, Preeti & Singh, Nipur. (2019). An Efficient Algorithm for Data Pre-Processing and Personalization in Web Usage Mining. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING. 7. 160-164. 10.26438/ijcse/v7i5.160164.

[12] HubSpot, Start Using Personalization Today: HubSpot Marketing Software, Start Using Personalization Today — HubSpot Marketing Software. [Online]. Available: https://www.hubspot.com/products/how-personalization-works. [Accessed: 13-Oct-2019].

[13] S. Mullin, Why Content Personalization Is Not Web Personalization (and What to Do About It), CXL, 21-Aug-2019. [Online]. Available: https://conversionxl.com/blog/web-personalization/. [Accessed: 13-Oct-2019].

[14] O. Smilansky, Navigating BIG DATA for BIG PROFITS. (Cover Story)., CRM Magazine, vol. 19, no. 10, Oct. 2015, pp. 2831 [Online]. Available: http://search.ebscohost.com.ezproxy.lewisu.edu/login.aspx?direct=true&db=a9h&AN=110155757&site=ehost-live&scope=site. [Accessed October 13, 2019]

[15] B. Raufi, F. Ismaili, J. Ajdari and X. Zenuni, Web personalization issues in big data and Semantic Web: challenges and opportunities., Turkish Journal of Electrical Engineering and Computer Sciences 27 (4): 237994. [Online]. Available: doi: 10.3906/elk-1812-25. [Accessed October 13, 2019]

[16] N. Behram and K. Crabtree, Big data, little ethics: confidentiality and consent, Forced Migration Review, no. 61, June 2019, pp. 46. [Online]. Available: http://search.ebscohost.com.ezproxy.lewisu.edu/login.aspx?direct=true&db=a9h&AN=137734617&site=ehost-live&scope=site. [Accessed October 13, 2019]

[17] J. Frankenfield, General Data Protection Regulation (GDPR), May 8, 2019 [Online]. Available: https://www.investopedia.com/terms/g/general-data-protection-regulation-gdpr.asp [Accessed October 13, 2019]

[18] E. Buchanan, Considering the Ethics of Big Data Research: A Case of Twitter and ISIS/ISIL. PLoS ONE, vol. 12, no. 12, Dec. 2017, pp. 16. [Online]. Available: doi:10.1371/journal.pone.0187155. [Accessed October 13, 2019]

[19] A. Mari, THE RISE OF MACHINE LEARNING IN MARKETING, ResearchGate, 2019. [Online]. Available: https://www.researchgate.net/publication/332865857_The_Rise_of_Machine_Learning_in_Marketing_Goal_Process_and_Benefit_of_AI-Driven_Marketing. [Accessed: 13-Oct-2019].