Beyond Who and What: Data Driven Approaches for User Characterization

Aastha Nigam

Department of Computer Science and Engineering
Interdisciplinary Center for Network Science and Applications (iCeNSA)
University of Notre Dame, Notre Dame, IN 46556
anigam@nd.edu

ABSTRACT

Social media and technology have drastically transformed the social and information networks around us. They have impacted how we communicate with others, search for information, and even how we express our personal opinions. Further, in this era of big data, not only are the online services collecting vast variety of user data, but we, as users, are also readily divulging significant amounts of information. Together, massive datasets obtained from diverse sources such as organizations and user generated content give us the opportunity to explore and understand complex behavior of both individuals and communities. This proposal aims at designing generalizable and scalable data-driven frameworks to gain a deeper understanding of the users, explain their actions and preferences, and infer personal traits. The proposed models will enable us to move beyond asking the conventional questions of who and what, and reveal answers about how and why. Given the varying digital persona of users motivated by their personal preferences and social attributes, we characterize users in two distinct domains: online health and peace studies. The models are designed to solve various real-world challenges to maximize their broader impact.

CCS CONCEPTS

• **Information systems** → **Data mining; Web searching and information discovery**; • **Computing methodologies** → **Machine learning**; • **Social and professional topics** → *User characteristics*; • **Applied computing** → *Law, social and behavioral sciences*;

KEYWORDS

user behavior; computation social science; user modeling; online healthcare; peace studies

ACM Reference Format:

Aastha Nigam. 2018. Beyond Who and What: Data Driven Approaches for User Characterization. In *WSDM 2018: 11th ACM International Conference on Web Search & Data Mining, February 5–9, 2018, Marina Del Rey, CA, USA*. ACM, New York, NY, USA, Article 4, 2 pages. <https://doi.org/10.1145/3159652.3170455>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5581-0/18/02...\$15.00

<https://doi.org/10.1145/3159652.3170455>

1 INTRODUCTION

With the increasing pervasive nature of technology, users today are leaving behind digital footprints and readily divulging significant amounts of information — giving rise to massive datasets with untapped potential to understand user behavior at a large scale. However, the behavior of each individual is unique and can be considered to be influenced by many confounding factors at a personal and social level — guiding their interests, preferences and consumption patterns. The nuances in these digital personas can be explained by the fundamental demographic, psychographic and geographic differences between users. Therefore, in this proposal, we focus on studying the behavior of different individuals in diverse domains and provide data-driven frameworks for understanding, modeling and predicting both individual behavior, as well as collective phenomena.

We aim to harness the power of data from diverse sources such as information collected by organizations, user generated content and census reports, to build a deeper understanding of online users, explain their actions and preferences, infer personal traits and potentially predict their next action. By modeling such diverse nuggets of data and integrating user attributes, we argue that we can move beyond asking traditional questions of “who” and “what”, and instead answer integral questions about user behavior such as “how” and “why” ([2]). Such an understanding can not only unveil individual behavior, but can also reveal macroscopic properties of communities. Specifically, given the complex and multi-faceted nature of the human behavior, we formalize a set of data science problems and propose computational models to build a comprehensive understanding of users in diverse domains such as online health and peace studies. Firstly, we focus on developing a profile for an online health seeker and examining the most informative features for understanding their health reading habits and preferences. Secondly, we propose a content driven framework to generate synthetic user profiles for user attribute inference in case of imbalanced datasets. Thirdly, we evaluate opinion and sentiment for each individual and measure the collective impact of the masses on socio-political events and finally we present a growth model to capture and predict activity trends of polarized users.

The proposed work demonstrates the significance of personal attributes such as age and gender, and social attributes, in defining the attitude, preferences, choices and behavior of a user. It also highlights diversity in information consumption/generation depending on the role an individual seeks — whether it be an online health consumer or a political tweeter. Further, the data driven frameworks designed and implemented in the research use computational models such as factorization machines and deep learning

methods, which provide an effective and efficient way to characterize user behavior. Finally, the proposed problems and solutions give rise to important implications for real-world applications in user behavior understanding.

2 PROPOSED WORK

In this section, we discuss problem statements from varied user domains, present proposed data driven frameworks, and describe the results obtained for each project.

2.1 User Understanding in Online Healthcare

Various online sources including specialized websites such as WebMD¹ and Everyday Health² are commonly used by individuals to answer health queries about themselves. According to previous studies, many confounding factors can influence the health of an individual including demographic, socio-economic, cultural and environmental factors, which subsequently impact the health content a user would be interested in [3]. Given this close association between the factors impacting health of an individual and their subsequent health seeking behavior, we study the online health consumption of users for a large U.S. based online healthcare provider.

We firstly characterize online health consumption of users across contextual dimensions such as demographics, time and socioeconomics. Our analysis demonstrates that health interests vary substantially based on the age, gender and location attributes of an online health seeker. Further, we build a supervised model using factorization based methods to explain the topical preferences of an individual motivated by such contextual dimensions. Our analysis highlights the importance of using demographic features to understand and infer the health preferences.

2.2 Synthetic User Generation for Imbalanced Data

Given the large volumes of content being consumed by each user, the aim of this research was to predict the gender of the users given the articles they consume. Unfortunately, many real world datasets suffer from an imbalance in their user base with respect to varied attributes such as age and gender. Therefore, to address the challenge of imbalance and to infer the user attribute based on the consumed content, we proposed the URI (*User Representation and Inference*) framework. The framework builds a representation of users based on the content they have consumed. It consists of four main components: 1) Document vectorization, 2) Generation of user content profiles, 3) Synthetic user generation and lastly, 4) User attribute inference.

After learning representative profiles of users based on the content they have consumed, we address imbalance by building synthetic profiles for the minority user class. We use generative adversarial networks (GANs) [4] to learn a distribution of existing users. Next, we sample synthetic instances from this distribution to over-sample the minority class and create a more balanced dataset. Finally, we leverage this dataset to build a supervised model to infer the gender of the users [5].

¹<https://www.webmd.com/>

²<https://www.everydayhealth.com/>

2.3 Social Signals to Inform Peace Processes Implementation and Monitoring

In the last few years, peace processes have become the dominant outcome of civil war terminations. While they have become the preferred outcome, they are protracted and contentious processes which require a significant amount of compromise and bargaining from all participants. Due to this complexity, they are subjected to a 40% chance of failure in the first 10 years of their implementation. Further, given their bureaucratic nature, they are often monitored using government and newspaper reports. As a consequence, while the peace processes are designed to bring peace to the masses, they usually lack the pulse of the nation. Therefore, we propose the use of social media to harvest key indicators such as inter-group polarization and public sentiment to study their implementation and monitor their progress.

Using 2016 Colombia final peace agreement and national referendum, we discover the Yes group (in favor) and the No group (against) and characterize the polarized groups based on their activity and content. Further, we build a predictive framework to infer the sentiment of the tweets leveraging tweet structure, content and user network features. Finally, we compare results obtained from social media analysis to compare with actual polling and referendum results. We demonstrate how social media provides large volumes of information that can be leveraged in conjunction with polls, to better understand the success or failure of a peace process. Beyond peace processes, this work also highlights how public opinion can be of key importance to other sociopolitical events such as military operations and immigration [6].

2.4 Growth Models for Polarity Change

While individuals in polarized populations behave differently, we aim to model the joint activity volume trends and predict their activity for time t_{n+1} given the trend from time $t_1, \dots, t_n$. In order to address this problem, we leverage the use of one of the most classic growth models in epidemiology — the Susceptible Infected (SI) model [1] as the foundation. We propose to not only leverage the growth of one group, but also study how the interaction between two differently polarized individuals affects the volume trends.

Acknowledgments: The Ph.D. study is supported in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 and the National Science Foundation (NSF) Grant IIS-1447795.

REFERENCES

- [1] Roy M Anderson, Robert M May, and B Anderson. 1992. *Infectious diseases of humans: dynamics and control*. Vol. 28. Wiley Online Library.
- [2] Alex Beutel. 2016. User Behavior Modeling with Large-Scale Graph Analysis. *Ph.D. Thesis* (2016).
- [3] Göran Dahlgren and Margaret Whitehead. 1991. Policies and strategies to promote social equity in health. *Stockholm: Institute for future studies* (1991).
- [4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
- [5] Aastha Nigam and Nitesh V Chawla. 2017. A Content-Driven Framework for Online User Understanding. (2017).
- [6] Aastha Nigam, Henry K Dambanemuya, Madhav Joshi, and Nitesh V Chawla. 2017. Harvesting social signals to inform peace processes implementation and monitoring. *Big data* 5, 4 (2017), 1–19.