Quantifying Cross-platform Engagement through Large-scale User Alignment

Jiejun Xu, Tsai-Ching Lu, Ryan Compton, David Allen HRL Laboratories 3011 Malibu Canyon Road Malibu, CA 90265 {jxu, tlu, rfcompton, dallen}@hrl.com

ABSTRACT

As online social media becomes prevalent as part of our daily life, it is increasingly common for a user to have accounts on multiple social media platforms. In this work, we present our findings on quantifying the extent of user engagement of different platforms as well as their correlations. The study is conducted based on a large-scale user alignment on 6 major social media platforms. Specifically, we identify both explicit and implicit mentions of social media accounts from the Twitter Decahose stream over a period of 22 months. During the process, we have aligned a total of 21, 456, 808 Twitter users to their alternative accounts on different platforms. Subsequently, we extract the number of overlapping users between any combination of these social media platforms exhaustively.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous; J.4 [Computer Applications]: Social And Behavioral Sciences

General Terms

Measurement, Human Behavior

1. INTRODUCTION

Online social media has gained tremendous popularity over the past decade. Each social media platform has different characteristic and usually focuses on different aspect of the user needs. For instance, Twitter and Facebook focus more on communication, while Tumblr and Instagram focus more on photo sharing. As a results, it is common for a user to have multiple accounts on different social media platforms to accommodate his needs. In this work, we present a study on quantifying the level of user engagement on different combinations of the social media platforms. Specifically, we conduct a large-scale user alignment across 6 major social media platforms to extract the number of overlapping

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). Copyright is held by the author/owner(s).

WebSci'14, June 23-26, 2014, Bloomington, IN, USA.

ACM 978-1-4503-2622-3/14/06.

http://dx.doi.org/10.1145/2615569.2615662.

users. These platforms include Twitter, Tumblr, Wordpress, Blogger, Instagram, and Facebook.

Several recent works have started to investigate the crossplatform alignment problem for user modeling, personalization and recommendation. For instance, Abisheva et al. [1] combined user data from Twitter and YouTube to provide a descriptive analysis of the demographics and behavioral features on the two online communities. Vu et al. [7] proposed a general and extensive system architecture for aggregating and integrating users' social profiles into collaborative systems. Tiroshi et al. [6] modeled online users based on features extracted from different social media sites. They showed that aggregated user profile enhanced personalization service effectively. Similar study has been conducted by Deng et al. [2] to improve personalized video recommendation. There are also studies focused on developing methodologies to accurately map users across platforms. Zafarani et al. [9] introduced a behavioral-modeling based approach to connect individual across social media sites. Other recent works on user mapping utilized network attributes [3] and a variety of content features [5] to de-anonymize users across sites.

Most prior work focused on aligning users with relatively confined datasets. In contrast, the main contribution of our effort is the cross-platform user alignment *at scale*, using a large real-world data stream to derive representative results for major social media platforms. To the best of our knowledge, this is the first attempt to quantify overlapping social media usage from a variety of platforms in such a scale.

2. USER ALIGNMENT

We take a simple extractive-based approach to identify mentions of social media user accounts from the Twitter datastream, which consists of 10% sample of public tweets

Tumblr	http://[www.]*[a-zA-Z0-9]+.tumblr.com http://tmblr.co/(\\S{4,20})\	
Wordpress	http://[www.]*[a-zA-Z0-9]+.wordpress.com http://.*wp[.]me.*	
Blogger	http://[www.]*[a-zA-Z0-9]+.blogspot.com N/A	
Instagram	$\label{limit} $$ $$ $ \begin{array}{ll} $ & \text{http://[www.]*instagram.com/[a-zA-Z0-9]+} \\ & \text{http://instagr} \\ & \begin{array}{ll} & \begin{array}{ll} & & \\ & & \end{array} \\ & \begin{array}{ll} & & \\ & & \end{array} \\ & \begin{array}{ll} & & \\ & & \end{array} \\ & \begin{array}{ll} & & \\ & & \end{array} \\ \end{array}$	
Facebook	ebook http://www.facebook.com/(\\S+) N/A	

Figure 1: Regex patterns used for identifying social media accounts from the Twitter data. Note that the matched Short URLs are resolved to full account address in post processing.

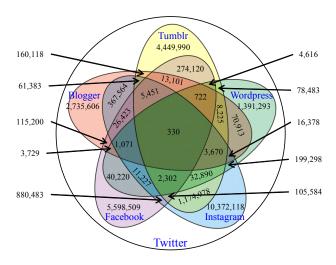


Figure 2: A Venn Diagram showing the number of aligned users from different combinations of social media platforms.

from April 2012 to January 2014 obtained through the GNIP Decahose. The full dataset amounted to 67.2TB of uncompressed JSON records. Each record contains the content of the tweet as well as the profile information of the corresponding Twitter user. Two types of user account mentions are searched in our work. The first type is the Explicit Self-reported accounts. We search for every user profiles in our Twitter corpus with the regular expressions, which indicate different social media platforms. For instance, we searched for the pattern "http://[www.]*[a-zA-Z0-9-_].tumblr.com" for mentions of Tumblr user accounts. The second type of account mention is obtained through Implicit Cross-Links. Many existing social media sites support content synchronization in order to reduce end user effort. This allows a user to submit a post from one platform, and publish the same content to all other social media accounts under him or herself. For the case of Twitter, there is usually a URL appended at the end of a tweet to indicate its origin. For example, if a post was original published in Tumblr, and the content was synchronized to Twitter. There would be a short URL encoded at the end of the tweet referring back to the Tumblr post (e.g., http://tmblr.co/ZVxw1y15H_Go3). The original Tumblr user account can be obtained by resolving the short URL. This type of cross-referencing is very useful in terms of identifying the same user across the two platforms. This is because the short URL (with prefix pattern "tmblr.co") is automatically generated by the Tumblr server, and it would be triggered by the synchronization process between two linked accounts. Since synchronization only happens when a user owns both accounts, this is a reliable way to identify same users across platforms. A more detailed evaluation is presented in [8] to demonstrate the effectiveness and accuracy of this approach. Overall, this method results in 95% accuracy in terms of linking user across platforms. The rare cases of error are due to direct "copy-and-paste" of tweet messages. Figure 1 summarizes the Regex patterns used for both explicit and implicit identification of account mentions of different platforms. Note that we do not search the implicit cross-link patterns for Blogger and Facebook, as the short URLs generated from both platforms can't be resolved to user accounts exclusively.

	Total Aligned Pairs	Unique Users (By Twitter ID)	Geo-coded Users
Tumblr	6,549,937	4,449,990	3,593,058
Wordpress	1,766,825	1,391,293	887,986
Blogger	3,042,319	2,735,606	1,794,895
Facebook	6,545,446	5,598,509	4,040,912
Instagram	10,372,118	10,372,118	9,605,944

Figure 3: User alignment and geocoding statistics derived from Twitter.

DISCUSSION 3.

A total of 21, 456, 808 Twitter users have been aligned to other accounts in various social media platforms. Subsequently, we compute the number of overlapping users for different combination of social media platforms exhaustively. The Venn diagram in Figure 2 summarizes our findings. New platforms such as Instagram, Tumblr and Facebook appear to have more penetrations on the alignment, while older blogging platforms such as Blogger and Wordpress seem to have less. User alignment allows for information reuse across platforms. For instance, user geolocations are derived in [4] for the Twitter platform, and a substantial amount of this information propagates to other platforms through alignment (see Figure 3). Our future work may include user profile modeling, interest prediction, and collaborative summarization from aligned accounts in multiple social media platforms.

 $\textbf{4.} \underbrace{\textbf{ACKNOWLEDGMENTS}}_{\text{Supported by the Intelligence Advanced Research Projects Activity}}$ (IARPA) via Department of Interior National Business Center (DoI / NBC) Contract Number D12PC00285. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the author(s) and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

- **REFERENCES**[1] A. Abisheva, V. R. K. Garimella, D. Garcia, and I. Weber. Who watches (and shares) what on youtube? and when? using twitter to understand youtube viewership. CoRR, 2013
- [2] Z. Deng, J. Sang, and C. Xu. Personalized video recommendation based on cross-platform user modeling. In International Conference on Multimedia and Expo, 2013.
- [3] P. Jain, P. Kumaraguru, and A. Joshi. @i seek 'fb.me': Identifying users across multiple online social networks. In Proceedings of the 22Nd International Conference on World Wide Web Companion, WWW '13 Companion, 2013.
- [4] D. Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In International Conference on Weblogs and Social Media, 2013.
- M. Korayem and D. J. Crandall. De-anonymizing users across heterogeneous social computing platforms. In International Conference on Weblogs and Social Media, 2013.
- A. Tiroshi, S. Berkovsky, M. A. Kaafar, T. Chen, and T. Kuflik. Cross social networks interests predictions based ongraph features. In Proceedings of the 7th ACM Conference on Recommender Systems, RecSys '13, 2013.
- [7] X.-T. Vu, M.-H. Abel, and P. Morizet-Mahoudeaux. Integrating social network data for empowering collaborative systems. Advances in Intelligent Systems and Computing. Springer, 2013.
- J. Xu, R. Compton, T.-C. Lu, and D. Allen. Rolling through tumblr: Characterizing behavioral patterns of the microblogging platform. In ACM Web Science Conference, 2014.
- [9] R. Zafarani and H. Liu. Connecting users across social media sites: A behavioral-modeling approach. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, 2013.