Modeling the Dynamics of User Content Generation: A Case Study of Stack Exchange Websites

Author Name
Department
Institution
Email

ACM Reference format:
Author Name. 2017. Modeling the Dynamics of User Content Generation: A Case Study of Stack Exchange Websites. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 3 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Ignore the Introduction for now. Content generation and subsequent consumption is a key feature of the social web experience. Every second, on average, several thousand tweets are tweeted on Twitter, which corresponds to several hundred million tweets per day and several hundred billion tweets per year. The content generation stats are in similar scale for other massive social media platforms such as Facebook and YouTube. We continue to witness the Web 2.0 era in the light of these user-generated content (UGC) platforms, where content generation is instrumental in preserving sustainability of the platforms– serving 'nowness' in the social web experience.

While content generation is at the heart of sustainable UGC platforms, our understanding of what factors are effective at encouraging content generation and associated growth is yet limited. To date, growth of UGC platforms has been studied mainly from the perspective of population dynamics, with researchers emphasizing interacting population based models to explain user growth in these platforms. These models, however, only consider the interaction among users, and ignores the interaction among users and content– consequently, can not explain content growth. There is a growing body of empirical studies that analyze patterns of user content generation– primarily temporal patterns– revealing useful insights. However, even with these insights, there is a gap in realizing the factors of content growth and corresponding dependency. What is missing from the picture are models of content growth that clearly establish the relationship between content growth and associated factors. Developing such models is important for understanding content growth dynamics, predicting future growth, and identifying platforms that will sustain for a long period of time.

Conference'17, July 2017, Washington, DC, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 MODELING USER CONTENT GENERATION

In this section, we first describe the desired properties of our user content generation model, and briefly discuss some unsuccessful model attempts. We then introduce our factor based user content generation model for CQA platforms.

2.1 Desiderata

To better understand the dynamics of user content generation in CQA platforms, we opt for a model with the following desired properties.

D1. *Macro-scale:* The model should capture user content generation in CQA at aggregate level.

D2. *Explanatory:* The model should give us some deeper understanding of how the aggregate user community generates content.

D3. *Predictive:* The model should allow us to make predictions about future content generation and relevant aspects.

D4. *Minimalistic:* The model should have as few parameters as necessary, and still closely reflect reality.

D5. *Comprehensive:* The model should encompass the dynamics of content generation for different types of content (e.g., question, answer, comment) in varieties of CQA platforms.

2.2 Some Unsuccessful Attempts

We considered several alternative models to comprehend the content generation dynamics in CQA platforms. Our first attempt was to model user content generation as 'self-exciting' point process. In this attempt, we designed and implemented several variants of Hawkes process. Our second attempt was to model user content generation using 'stage-structured' projection matrix. In this attempt, we built variants of Leftkovitch matrix. Although both variants of models showed promise, they failed to meet one or more requirements mentioned above. In particular, the Hawkes process based models lacked long term prediction accuracy, while the Leftkovitch matrix based models lacked prediction accuracy and minimalism (large number of parameters).

2.3 Factor Based Model

Motivated by the macroeconomic production models, we focused on designing factor based model for user content generation in CQA platforms. Instead of directly modeling user content generation as a dynamic process (function of time), we model it in terms of associated factors, which themselves are dynamic. From this point forward, we report the factors of content generation for different content types; discuss potential basis functions to capture the effect of a single factor, and the potential interaction among multiple factors; and finally introduce the alternative models based on the basis functions and factor interactions.

Factors of Content Generation. We recognize two key factors that drive content generation in CQA platforms, namely user participation and content dependency.

User Participation: The number of active users is the most important factor in deciding the number of generated content. The participation of more users induce more questions, answers, and comments in a CQA platform.

Content Dependency: While user participation is vital for content generation, content dependency also affects the number of generated content for different types. Content dependency implies the dependency of a type of content on other type of content(s). For example, answer generation relies on question generation– in absence of questions, there will be no answers.

Symbol	Definition
$U_q(t)$	# of users who asked questions at time t
$U_a(t)$	# of users who answered questions at time t
$U_c(t)$	# of users who made comments at time t
$N_q(t)$	# of active questions at time t
$N_a(t)$	# of answers to active questions at time t
$N_c^q(t)$	# of comments to active questions at time t
$N_c^a(t)$	# of comments to active answers at time t
$N_c(t)$	# of comments to active questions/answers at time t
f_w	The functional relationship for content w

Table 1: Notations used in the models

Based on the above discussion, we identify the key factors of content generation in Stack Exchange websites for three primary content types: question, answer, and comment. Figure 1 shows these factors using notations from Table 1. These factors lead to the following functional relationships in all Stack Exchange websites.

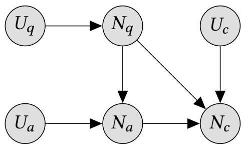

Figure 1: Factors of content generation in Stack Exchange

- There is a single factor in generating questions: users who ask questions (aka askers).

$$N_q = f_q(U_q)$$

- There are two key factors in generating answers: questions, and users who answer questions (aka answeres).

$$N_a = f_a(N_q, U_a)$$

- There are three key factors in generating comments: questions, answers, and users who make comments on these questions and answers (aka commenters).

$$N_c^q = f_{c^q}(N_q, U_c)$$

$$N_c^a = f_{c^a}(N_a, U_c)$$

$$N_c = N_c^q + N_c^a$$

The aforementioned functional relationships imply that the dynamics of user content generation for each content type depends on the function describing factor dependent content generation, and the availability of factors. These equations embody two critical assumptions. First, they assume that different content types interact only through their use of factors— how each content type consumes a factor. Secondly, they assume that functions describing content generation depends on the interaction among the factors– how the factors are related.

Basis Functions. We use basis functions to capture the effect of a given factor on a particular type of content. While there is a variety of basis functions available for regression, we consider three basis functions widely used in economics and growth modeling, namely power– $g(x) = ax^{\lambda}$, exponential– $g(x) = ab^{x}$, and sigmoid– $g(x) = \frac{L}{1+e^{k(x-x_0)}}$.

Interaction among the Factors. We use

REFERENCES