Identifying impact of intrinsic factors on topic preferences in online social media: A nonparametric hierarchical Bayesian approach

Yezheng Liu [a,b], Jiajia Wang [a,d], Yuanchun Jiang [a,b,*], Jianshan Sun [a,b], Jennifer Shang [c]

[a] *School of Management, Hefei University of Technology, Hefei, Anhui 230009, China*
[b] *Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei, Anhui 230009, China*
[c] *The Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, USA*
[d] *BigData PDU, Huawei Software Technologies CO.LTD, Nanjing, Jiangsu 210012, China*

ARTICLE INFO

ABSTRACT

Social media offers a new communication channel for users and affords an interactive opportunity between users and the firms about the products and the brands. Understanding what topics are important to users and the corresponding internal motivation is crucial for managers to successfully engage customers and promote business through social media. Assuming topic preference is the outcome of intrinsic factors such as gender, age and personality traits, this paper proposes an improved nonparametric hierarchical Bayesian topic (NHBT) model to investigate the multiple-to-multiple generative relationships from intrinsic factors to topic preferences. The proposed NHBT model employs a three-level generation framework based on Dirichlet process to study the impact of intrinsic factors on users topic preference. Our study of Facebook data shows that NHBT model is able to draw valuable latent topics (e.g. music band, chemical biology, cosplay) from the open social media environment, and reveal the internal motivation for users topic selection behaviors (e.g. users with low conscientiousness and high extraversion personality prefer topics about campus party). Our experiments also show that NHBT model can identify the intrinsic factors dominating topic preferences for individual users, and provide foundations to predict the intrinsic factors for new user generated contents.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

With the development of the social media and the fast growing number of users, online social media has become the best platform for users to communicate with each other. In 2014, the number of social media users has reached 1.87 billion and it is estimated to approximate 2.55 billion in 2018. Through online social media platform, users make friends, share interests and communicate their concerns about firms, products and brands. Online social media has been an important channel for firms to attract new customers, improve brand perception and grow their business.

---

* Corresponding author.
E-mail addresses: liuyezheng@hfut.edu.cn (Y. Liu), wjj5881005@126.com (J. Wang), ycjiang@hfut.edu.cn (Y. Jiang), sunjs9413@hfut.edu.cn (J. Sun), shang@katz.pitt.edu (J. Shang).

Because of the increasing popularity and the unprecedented volume of data available in online social media, what topics are of users interest has been an important concern for firms. Various generative probabilistic models such as LDA[2], sLDA [21] and L-LDA [29] have been proposed to summarize users topic preferences from contents published in online social media. While these models are able to predict what topics users prefer[12,14,28], little works have been done to study why users prefer the topics. Cao argued that it is ineffective or even impossible to deeply scrutinize users native behavior intention if we employ entirely web usage data to analyze user behavior [3]. Understanding the incentive reason of users online social behavior provides managers a better perspective to promote their business. It helps firms gain insight into the decisions of online social networking groups [38], increase recommendation diversity and accuracy [6], and enhance customers online purchase intention [45].

Intrinsic factors are the internal motivations for a persons actions and are widely used to explain users social networking behaviors, such as the online language usage[25,34], emotion expression [1,20], advertisement preference [16], and music selection [10]. The literature proves the correlation between the internal motivations and online social behaviors. However, they tend to focus on the correlation test, with little emphasize on the generative relationships from the intrinsic factors to extrinsic behaviors. The psychology literature maintains that extrinsic behaviors are caused by intrinsic factors such as gender, age and personality [13]. The correlation tested in the current literature does not reveal the internal motivation of user behaviors. When using social media data to model internal motivation, the current literature focuses on closed-vocabulary methods without emphasizing diversified topics in online social media. They often classify user generated contents into 64 categories (e.g. topics) by the Linguistic Inquiry and Word Count (LIWC) software, and then explore the correlation between the intrinsic factors and the fixed LIWC categories [22,27,34]. The topics discussed evolve over time and free-form languages are widely adopted. The closed-vocabulary embedded in the LIWC software obviously cannot capture the topic dynamics and may cause information loss.

In this research, we propose a hierarchical Bayesian topic model to investigate the impact of intrinsic factors on users topic preferences. Using the open-vocabulary rather than closed-vocabulary, the proposed model avoids the need of predefined topic number and types, and draws latent topics dynamically from online social media. It is widely documented that the influence of personality on users social networking behavior differs by both gender and age [7,23]. A specific dimension of personality is also reported to have influence on different social networking behaviors. For example, agreeableness and conscientious are positively related to the religion topic [34], and negatively related to negative emotion words [44]. Therefore, different from the existing models which associate user behavior with one intrinsic factor at a time, we assume a multiple-to-multiple generative relationship from intrinsic factors to topic selection behaviors. That is, a specific topic selection behavior is the outcome of multiple intrinsic factors [34,44] and an intrinsic factor may have influence on multiple topic selection behaviors. Because it is too restrictive and not feasible to fix the number of intrinsic factors generating a specific behavior [30,37], we employ a nonparametric approach to model the process of generating topics from intrinsic factors.

The proposed nonparametric hierarchical Bayesian topic (NHBT) model employs a three-level generation process to investigate the internal motivation of topic selection behaviors. The first level (global level) defines the global random measure, which is a draw from a standard Dirichlet process [37]. The second level (factor level) defines random measures for intrinsic factors, which are draws from a Dirichlet process, where the base measure is the random measure in global level. Finally, the third level (document level) defines random measures for each document. The base measure of the Dirichlet process in document level is a mixture of random measures in the factor level. The three-level process is iterated and we propose an inference approach under the framework of Gibbs sampling. The numerical studies based on data collected from myPersonality application, a third-party application on Facebook, show that the proposed model can find diverse topics and reveal the internal motivation regulating users topic selection behavior. The contributions of the proposed model are threefold.

(1) To the best of our knowledge, this is the first paper to investigate the generative relationships between intrinsic factors and user-generated contents based on nonparametric hierarchical Bayesian topic framework. The proposed NHBT model is consistent with the action logic of behavior and internal drives, and can capture the dynamic character of online social media.

(2) Conventionally, the relationship between intrinsic driver and user behavior is regarded as a one-to-one mapping. Our model more realistically proposes a three-level derivation process which generates each topic from multiple intrinsic factors and also correlates each intrinsic factor with more than one topic. The proposed NHBT model allows statistical strength to be shared across intrinsic factors, topics and documents.

(3) Instead of applying LIWC software to obtain the pre-defined topics, the proposed NHBT model automatically extracts latent topics from online social media. Without the restriction of the closed LIWC dictionary, the proposed model is able to obtain more meaningful topics.

The remainder of the paper is organized as follows. Section 2 surveys the related work on the automatic behavior analysis of personality and topic models for labeled data. We detail the proposed NHBT model and the inference process in Section 3. Both the quantitative and qualitative evaluation results of the proposed model are presented in Section 4. Section 5 gives the conclusions and the directions of future research.