Internal/External Information Access and Information Diffusion in Social Media

Tian Xia[1], Xing Yu[2], Zheng Gao[3], Yijun Gu[4], Xiaozhong Liu[3]

[1]School of Information Resource Management, Renmin University of China

[2] Dept. of Human-Centered Computing, Indiana University Indianapolis

[3] Dept. of Information and Library Science, Indiana University Bloomington

[4] School of Cyber Security, People's Public Security University of China

Abstract

As social media platform not only provide infrastructure but also actively perform algorithmic curation for profit and user experience, it leads to an information filter bubble phenomenon: users are trapped in their own personalized bubble and are exposed only to the opinions that conform their beliefs and interests, thus potentially creating social polarization and information islands. However, filter bubbles hardly restrict all the users in a large social network, some information explorers can break the bubble and bring external global knowledge back to the internal network. In this paper, we investigate this assumption via hashtag adoption prediction. First, we construct a heterogeneous graph and extract 17 features to describe the event of hashtag adoption. Then, we generate learning instances and train a lasso regression model to do prediction. Preliminary results show that information explorers are more likely to adopt new hashtags than others, thereby more internal and external information can be diffused via these special users.

Keywords: Social Media; Information Adoption; Information Diffusion; Information Access; Weibo

Citation: Xia, T., Yu, X., Gao, Z., Gu, Y. & Liu, X. (2017). Internal/External Information Access and Information Diffusion in Social Media. In iConference Proceedings 2017, Vol. 2 (pp. 129-133). https://doi.org/10.9776/17235

Contact: xiat@ruc.edu.cn

1 Introduction

The proliferation of social media is bringing about significant changes in how people perceive and make sense of their world (Pak & Paroubek, 2010; Shuai, Liu, Xia, Wu, & Guo, 2014). Millions of individuals communicate with each other through a variety of social media platforms, sharing pertinent information about the world as well as the most minute details of their social lives, thereby collectively shaping each others' culture and worldview. However, in addition to allowing information to travel freely through social ties, many social media platforms perform an information language/policy/network/algorithmic barrier. This curation raises an important concern, often referred "filter bubbles," where people are increasingly trapped in their own information "bubble"—being exposed only to information that conforms to their existing beliefs and political positions, potentially creating information "islands" and potentially social polarization. Note that, in a large social network, in most cases, filter bubbles hardly restrict all the users, and some information explorers can always break the bubble, while bringing some global knowledge back to the local network.

In this study, we investigate this interesting problem by leveraging massive Weibo data. Unlike most popular microblogging systems, most Weibo users (in China) are restricted in a local information network because of different local law reasons (Zhu, Phipps, Pridgen, Crandall, & Wallach, 2013). When Weibo users trying to access global information liek Facebook or Youtube, they usually have two alternatives, 1. using VPN or proxy servers; 2. access Weibo service outside mainland China. We define them as "Information Explorers", and call other Weibo users as common users. In this paper, we will study whether information explorers, who have global information access, tend to adopt and broadcast topics in social network, thereby resulting in some global knowledge back to the local network.

2 Related Work

Information diffusion has been widely studied in recent years, it can be defined as the process by which a piece of information (knowledge) is spread and reaches individuals through interactions (Zafarani, Abbasi, &

Liu, 2014). So far, a lot of effort have been made to modeling how information spreads in social network, and until recently, some researchers put more attentions to the filter bubble problems, and propose new method like cross social media recommendation to break the bubbles (Liu, Xia, Yu, Guo, & Sun, 2016; Liu, Yu, Gao, Xia, & Bollen, 2016).

Because of technique, culture and different country policy reasons, Twitter and Weibo, the most popular microblogging systems, are isolated from each other. Therefore, current research work about filter bubble isolation problem focused on comparing the difference of Twitter and Weibo, and try to find some way to connect these two social networks together(Shuai et al., 2014; Liu, Xia, et al., 2016). Besides above efforts for fighting with the filter bubbles, here we first propose another assumption that filter bubbles can be broken by some special users, who can access external information and bring it back to the internal network, and we verify this assumption by topic adoption prediction method.

3 Methodology

Are Information Explorers more likely to adopt and broadcast new topics in social media? To answer this question, we extract comprehensive features from Weibo retweet diffusion graph to predict information adoption behavior, and investigate whether the key features are strong related to these special users or not. If that's true, we can say that information explorers are more important for information access and adoption, thereby, they can introduce external information to the internal network. To verify this assumption, we do analysis via the following steps as shown in figure 1.

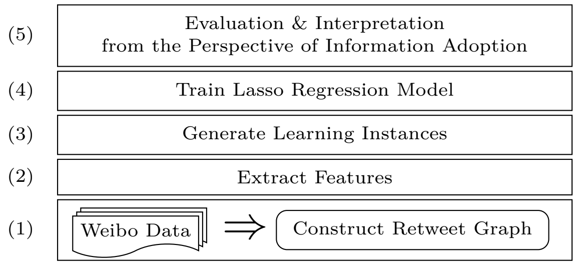
(5)
Evaluation & Interpretation
from the Perspective of Information Adoption
(4)
Train Lasso Regression Model
(3)
Generate Learning Instances
(2)
Extract Features
(1)
Weibo Data
Construct Retweet Graph

Figure 1: Analyzing Workflow.

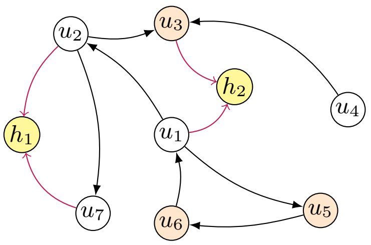

Figure 2: Heterogeneous Graph Example. (In this figure, u_i and h_j denote user and hashtag respectively, u_3, u_5 and u_6 with light pink background color are information explorers.)

First, we construct a heterogeneous graph to describe how users retweet messages and use hashtags. Hashtag can be treated as a topic, when a user introduces a new hashtag in his or her message, we say that the user adopt a topic. Formally, we define a topic adoption event as a user u_i, adopts a topic h_j, at time t(Liu, Yu, et al., 2016), denoted as a triple $event_t = (u_i, h_j, t)$.

Let $G = (\mathcal{V}, \mathcal{E})$ represents the graph for user-retweeting and hashtag adoption hybrid network, where each node $v \in \mathcal{V}$ represents a Weibo user u or a hashtag h, and edge $(u_i \rightarrow u_j) \in \mathcal{E}$ represents user u_i has retweeted a message posted by u_j, edge $(u_i \rightarrow h_j) \in \mathcal{E}$ represents user u_i has used hashtag h_j in the past. After G was created, all isolated nodes are removed to reduce the time and space cost. Figure 2 is an snippet of graph G, where h_1 and h_2 are hashtag nodes, and $u_i (i \in [1, 7])$ represents the user node, in particularly, u_3, u_5 and u_6 are information explorers in this figure.

The topic adoption model predicts if user u_i will adopt hashtag h_j in the future. Previous studies show that some features extracted from retweeting network are very useful for prediction (Yang, Sun, Zhang, & Mei, 2012), include: in-degree/out-degree, hashtag numbers that u_i used and prestige like PageRank score. In our study, we extract more features related with information explorers from above graph G, all the features we used are listed in table 1.

where f_i in table 1 is the i^{th} feature we extracted, u_i represents any user node of graph G, and h_j represents a hashtag. All neighboring explorers of current node u_i is denoted as set N_1, while N_2 is the set of neighboring explorers of all u_i's neighbors. Take user node u_1 in figure 2 as an example, its neighboring explorers set $N_1(u_1) = \{u_5\}$, and $N_2 = N_1(u_2) \cup N_1(u_5) = \{u_3, u_6\}$.

Given two time periods $T = [t_1, t_2]$ and $\Delta T = [t_2, t_3]$, we first construct graph G by using the data of period T, and then extract all features for each user from the graph G, these features are assumed as

Id	Feature Description	Id	Id	Feature Description
f_1	In-degree of u_i	f_6	f_{12}	The size of N_1(for f_6), N_2 (for f_{12})
f_2	Out-degree of u_i	f_7	f_{13}	Average PageRank of N_1(for f_7), N_2 (for f_{13})
f_3	PageRank of u_i	f_8	f_{14}	Maximum PageRank of N_1(for f_8), N_2 (for f_{14})
f_4	Number of u_i use h_j	f_9	f_{15}	Average in-degree of N_1(for f_9), N_2 (for f_{15})
f_5	u_i is an info-explorer or not	f_{10}	f_{16}	Average out-degree of N_1(for f_{10}), N_2 (for f_{16})
–	—	f_{11}	f_{17}	Total number of $u \in N_1/u \in N_2$ use h_j

Table 1: Features for Predicting Information Adoption

latent variables. Next, we collect all (u_i, h_j) pairs which meet: u_i do not use hashtag h_j in the time period T and do adopt h_j in the period ΔT, these pairs constitute the positive instances. Otherwise, if u_i do not use h_j in both period T and ΔT, it means a negative instance. Therefore, the dependent variable is boolean: 1 indicates u_i adopt h_j and 0 means u_i does not.

Once we get the learning instances, we use the lasso method (Tibshirani, 1996) to train the classification model, based on which we interpret the relationship between the extracted features and the topic adoption prediction. The lasso approach allows us to carry out feature selection while training the model by adding a $l1$ penality to the loss function of the logistic regression. It can help us select a most effective subset of all the candidate features. The penalized loss function is defined as:

$$L = \sum_i (y_i - \sum_p \beta_p x_{ip})^2 + \lambda \sum_p ||\beta_p||_1$$

where x_{ip} denotes the p^{th} feature in the i^{th} datum, y_i is the value of corresponding response, and β_p denotes the regression coefficient of the p^{th} feature. The last part $\sum_p ||\beta_p||_1$ is the $l1$ penalty, and parameter λ controls the penalty strength.

By adding the $l1$ penalty, the important features will have high regression coefficients, and irrelevant/redundant features coefficient will be shrunk to zero. Therefore, the coefficients of explorer-related features would be larger if the explorers are more important in information adoption and diffusion.

4 Experiment

We extracted Weibo users, hashtags, and various kinds of relationships from 12,362,489 Weibo messages. The data covered the time period T from September 17, 2012 to September 23, 2012 (7 days) and ΔT from September 24 to 25, 2012. We find 50836 information explorers by their geography locations, and the final graph G contained 328,065 nodes and 783,811 edges. From above dataset, we extracted 4,502 positive instances, and randomly sampled the same number of negative instances, to make the learning instances balanced.

We split all 9,004 instances into two groups, 6,303 instances(70%) for training the model, and 2,701 instances(30%) for testing. We employed a 10-folds cross validation to tune and find the optimized parameter λ for the shrinkage penalty. And the experiment was implemented and carried out using the R language.

Prediction	false	true
false	1214	108
true	136	1243

Table 2: Confusion Matrix

Id	Coefficient	Id	Coefficient	Id	Coefficient
f_1	0.0670	f_6	-0.4136	f_{12}	.
f_2	-0.0006	f_7	-2.315E-5	f_{13}	.
f_3	-2.869E-5	f_8	4.715E-6	f_{14}	4.163E-8
f_4	0.0007	f_9	0.0167	f_{15}	-0.0108
f_5	0.9318	f_{10}	-0.0067	f_{16}	-0.0015
—	—	f_{11}	0.5305	f_{17}	.

Table 3: Features Coefficient List Obtained by Lasso Regression Model

The confusion matrix from our validation process was shown in table 2. For the final model, the prediction accuracy has achieved 90.97% in total with a sensitivity of 89.93% and a specificity of 92.01%.

For feature interpretation, we presented the coefficients of all features in table 3, the table cells filled by dot symbols represent irrelevant features, while the most important features were highlighted by yellow background color. We also presented a plot in figure 3 to describe the lasso model changes as the feature added gradually.

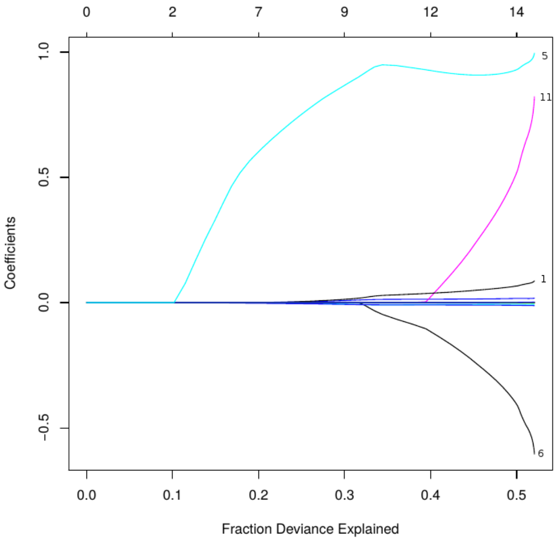

Figure 3: Changes of Lasso Regression Model When Features Are Added Gradually.

In figure 3, we presented how the coefficient of the each feature changes with deviance explained in the logistic regression model. This process can be better described as a process of adding predictors gradually into the model. In the figure, we saw a NULL model (to the left end) where no features were included and a full model (to the right end) where all the features were included. From the left end to the right end, we can see each feature was gradually added to the model. The slop of each feature represented the relationship between the feature's coefficient and the fraction deviance explained in the model. We noticed that at first the lasso result in a model contains only feature f_5. Then the rest of the feature entered the model gradually until all features were included. We found that feature f_5, feature f_{11}, and feature f_6 have steep slopes comparing with other features. In other words, these three features were considered to the most important features in the model. Their coeficients were shown in Table 3.

According to above evaluation results and interpretion, we draw the following points:

(1) Due to the strong positive correlation of feature f_5, information explorers have greater possibilities to adopt new hashtags, in another word, explorers were more likely to spread new topics in Weibo.

(2) The prediction performance has strong and positive correlation with the number of hashtags used by neighboring explorers, and negative correlation with the number of neighboring explorers. it means that user adoption behavior can be significantly affected by the user's neighboring explorers.

In summary, information explorers played an important role in topic adoption and diffusion, their information behaviors can help us to bridge the information gap between the inner social network and the outside.

5 Conclusion

In this study, we divide the Weibo users into two categories, i.e., common users and information explorers, and investigate the positive contributions of information explorers via predicting the topic adoption in Weibo social media. Based on feature analysis, we find that explorers have greater possibility than common users to adopt new hashtags, which means explorers tend to spread new information on Weibo. Therefore, they bring more information to others, and to a certain extent reduced the filter bubble problem .

We also find that neighboring explorers significantly affect the information adoption of users. If hashtag is exposed more times by the user's neighboring explorers, this user is more likely to adopt the hashtag later. However, if a user has more neighboring explorers, he/she has less possibility to use the hashtag.

One possible reason is that the user has already obtained some kind of information from the neighbors, and has low motivation to diffuse these redundant information.

References

Liu, X., Xia, T., Yu, Y., Guo, C., & Sun, Y. (2016). Cross social media recommendation. In *Tenth international aaai conference on web and social media.*

Liu, X., Yu, X., Gao, Z., Xia, T., & Bollen, J. (2016). Comparing community-based information adoption and diffusion across different microblogging sites. In *Acm hypertext.*

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *The international conference on language resources and evaluation.*

Shuai, X., Liu, X., Xia, T., Wu, Y., & Guo, C. (2014). Comparing the pulses of categorical hot events in twitter and weibo. In *Acm hypertext.*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012). We know what@ you# tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on world wide web* (pp. 261–270).

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: an introduction.* Cambridge University Press.

Zhu, T., Phipps, D., Pridgen, A., Crandall, J. R., & Wallach, D. S. (2013). The velocity of censorship: High-fidelity detection of microblog post deletions. In *Presented as part of the 22nd usenix security symposium (usenix security 13)* (pp. 227–240).