How Others Affect Your Twitter #hashtag Adoption? Examination of Community-based and Context-based Information Diffusion in Twitter

Chenwei Zhang[1], Zheng Gao[1], Xiaozhong Liu[1]
[1]Department of Information and Library Science, Indiana University Bloomington

Abstract
Twitter has become a rich source of people's opinions about a variety of topics, such as their daily life, and current news. Twitter's retweeting and mentioning mechanisms enable users to disseminate information broadly. In this study, we investigate the effects of community-based and context-based features on the users' information adoption and diffusion patterns in Twitter. Community-based features capture how the adoption of a hashtag by users within the target user's community and users outside that community influences the target user's selection of the target hashtag. Context-based features measure the influence of other users' adoption of hashtags that are semantically similar with a hashtag on the target user's adoption of this hashtag. We find the community-based features enhance the prediction of users' hashtag adoption and diffusion. However, the further exploration of context-based features is needed.

Keywords: Information diffusion; graph mining; community detection; context based similarity
doi: 10.9776/16538

Contact: zhang334@indiana.edu

1 Introduction

As the biggest microblogging platform in the world, Twitter is bringing about significant changes in how people perceive and make sense of their world (Simos, 2015). Prior studies showed Twitter has become a rich source of people's opinions about various topics (Pak & Paroubek, 2010). However, unlike other social media platforms, e.g., Facebook, no relationship reciprocation is required in Twitter, meaning a user can follow any other users, but don't need to be followed back by them (Kwak, Lee, & Park, 2010). This structure helps broadly disseminate information. The well-defined functionalities in Twitter also make the information adoption and diffusion easier for users. For example, the retweet mechanism "empowers users to spread information of their choice beyond the reach of the original tweet's followers" (Kwak et al, 2010, p. 591); the mention mechanism enables any of two users' conversation without any restriction; the hashtag sharing mechanism helps any users join discussion of certain topics freely.

Motived by the Twitter's novel features mentioned above, in this poster, we investigate the users' information adoption and diffusion patterns in this network. In Twitter, besides the underlying following/being-followed connections, users are interacting with others via various relationships, such as retweeting, mentioning, and sharing hashtags, which better reflect the information dissemination. The variety of user relations constructs a heterogeneous graph; each homogeneous graph, which only contains one type of user relations, can be separated. Specifically, we would like to know how the features embedded in the users' each network contribute to users' information adoption. It is obvious one user may have different positions in different homogeneous graph. Thus we further consider the community structures in each graph to examine whether community-based features better explain users' information adoption/diffusion behaviors. In addition to the influences through topological structures, sharing interests is also an important determinant for one user to adopt/diffusion a piece of information. So the semantic meanings of the information adopted, are taken into consideration and the context-based features are also tested for the capabilities to predict users' information adoption/diffusion.

Two hypotheses are verified in this study:

- Hypothesis 1 Community structures are useful in predicting users' information adoption/diffusion, i.e., using community-based features improves the prediction of users' information adoption/diffusion.
- Hypothesis 2 Can context-based features further improve the prediction of users' information adoption/diffusion?

2 Methodology

2.1 Data

In our experiment, we extracted all Twitter messages from September 17 to 25, 2012. Users, hashtags, retweeting, and mentioning information are included. There are 2,589,896 tweets in total. After reducing possible noises by removing inactive hashtags (used in fewer than ten messages) and inactive users (composing fewer than 5 messages), 797,869 users and 6,995 hashtags remain. Data of the first 7 days is used to collect the features (time T_1); the last 2-day is treated as time T_2 to verify the model and the features performance.

2.2 Homogeneous graph extraction and community detection

In Twitter network, we consider the following events: a user u_i adopts a new topic(hashtag) h_j, at time t; a heterogeneous network is generated by two relationships between users: a user mentions (m) the other user; a user retweets (rt) the other user. Thus two homogeneous graphs were extracted, $u \overset{m}{\rightarrow} u$; $u \overset{rt}{\rightarrow} u$, with isolated nodes removed.

We applied InfoMap algorithm to detect communities in each homogeneous graph, whose performance has been sufficiently proved (Lancichinetti & Fortunato, 2009). Each user belongs to only one community in each graph, but may have different communities in $u \overset{m}{\rightarrow} u$ and $u \overset{rt}{\rightarrow} u$.

2.3 Hashtag semantic meaning matching through Word2vec

We want to know the context when users adopt a certain hashtag, to better understand the topic. According to the Distributional Hypothesis (Firth, 1957), words having similar distributed properties tend to support similar meanings. For example, if there are two sentences like "*the citizens of X*" and "*the citizens of Y*", we know X and Y are semantically related since they occur in the same position. If *X* is a hashtag, extracting all its related terms helps understand the broad context of its topic, which is about "cities", rather than only treating the topic as the specific "*city X*". We applied the Word2vec tool (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) on the tweet messages to train a vector of semantically related terms for each hashtag to capture the context of information diffused in Twitter. Word2vec adopts a continuous bag-of-word model in this study and is useful for carrying semantic meanings.

2.4 Features extraction and information diffusion model construction

We collected the following features according to the features used in a study to model Twitter hashtag adoption by Yang, Sun, Zhang, and Mei (2012):

a) Indegree/Outdegree: Number of users a user u_i retweeting/being mentioned and number of users retweeting/mentioned by u_i.
b) Unique hashtags: Number of unique hashtags u_i used.
c) Popularity: Number/ratio of users adopting hashtag h_j.
d) Prestige: Prestige of hashtag h_j measured by the average/maximum PageRank score of all the users adopting it.

A logistic regression is used to model the features extracted in T_1 to predict each user u_i's adoption of hashtag h_j in T_2:

$$\Phi_x = \beta_{F_{x1}} F_{x1} + \beta_{F_{x2}} F_{x2} + \cdots + \beta_{F_{xk}} F_{xk} + \varepsilon$$

where x stands for the different homogeneous graph, F_{xk} the k th feature, $\beta_{F_{xk}}$ the coefficient demonstrating the prediction power of the kth feature, ε the intercept.

To verify Hypothesis 1, two regressions on each homogeneous network were conducted: one is the baseline, where the features were collected across the whole network; while in the other one, the values of features Popularity and Prestige were distinguished between intra-community, where the target user (whose information adoption is to be predicted) belongs to, and extra-community, where he/she does not belong to.

To verify Hypothesis 2, we added one more novel feature, users' content similarity with the topic, which captures how similar intra-community and extra-community members' tweets are with the hashtag to be adopted by the target user.

3 Preliminary results

Currently we sampled the top 50 highly used hashtags and investigate their adoption/diffusion. These most popular hashtags may trigger plenty of information diffusion. We generated the positive instances, in which users did not use the target hashtag in T_1, but adopted in T_2, and the negative instances, in which users did not use the target hashtag in T_1, and would not use in T_2. Since the number of negative instances is much larger than the positive ones, to ensure the quality of regression, we first picked-up all the positive instances from the set, and then randomly sampled the same number of negative instances. Finally 58,146 instances were used in fitting the regression models. Ten-fold cross validation was applied for evaluation. From the significant increase of corrected classified rate (CCR) and the decrease of root mean squared error (RMSE) in Table 1, we find the community-based features did enhance the prediction of users' hashtag adoption/diffusion. In addition, introducing community structures in user-retweeting-user homogenous graph gains a better performance than in user-mentioning-user graph (see highlighted part in Table1). Thus Hypothesis 1 is verified. The incorporation of context-based feature in the community-based model only improved the prediction in user-retweeting-user graph slightly; while no such change was observed in the user-mentioning-user graph. Hypothesis 2 is not fully verified and further efforts on this are needed.

	Correctly Classified Rate (CCR)		Root Mean Squared Error (RMSE)	
	$\overset{m}{u \to u}$	$\overset{rt}{u \to u}$	$\overset{m}{u \to u}$	$\overset{rt}{u \to u}$
Baseline	53.9975%	**54.7990%**	0.4949	**0.4938**
Community-based model	57.7805%	**62.8237%**	0.4843	**0.4708**
Community-combining-context-based model	57.4205%	63.5709%	0.4832	0.4663

Table 1. Evaluation results of three types of models

Table 2 shows detailed results of the community-based and community-combining-context-based models in both homogeneous graphs. By comparing coefficients of intra-community features in both graphs, we find them rather consistent (see highlighted part in Table 2). This implies the importance role of these intra-community features in predicting the information adoption/diffusion.

		Community-based		Community-based-combining-context-based	
Feature type	Feature	$\overset{m}{u \to u}$	$\overset{rt}{u \to u}$	$\overset{m}{u \to u}$	$\overset{rt}{u \to u}$
Intra-community features	Number of users for hashtag h	**9.65e-02**	**2.62e-02**	**7.30e-02**	**2.33e-02**
	Ratio of users for h	**1.38e+01**	**1.84e+01**	**1.09e+01**	**1.36e+01**
	Average users' prestige scores	**8.39e+03**	**-2.04e+04**	**9.57e+03**	**-2.14e+04**
	Maximum users' prestige scores	**-1.47e+03**	**-4.72e+02**	**-8.72e+02**	**-3.12e+02**
Intra-community-context feature	Average similarity score of users for hashtag h			-4.63e+01	-8.72e+00
Extra-community features	Average number of users for hashtag h	-5.10e+01	-8.59e+00	1.19e+02	7.38e+01
	Ratio of users for h	1.34e+02	6.64e+01	-7.91e+08	1.66e+09
	Average users' prestige scores	-6.92e+08	1.58e+09	3.03e+02	-7.03e+01
	Maximum users' prestige scores	6.60e+02	-6.01e+02	1.04e-02	9.15e-02
Intra-community-context feature	Average similarity score of users for hashtag h			4.88e-02	8.12e-03
Personal features	Target user u's indegree	-1.55e-03	8.10e-02	1.87e-01	2.09e-01
	u's outdegree	4.04e-02	-6.96e-05	2.65e+02	2.98e+02
	Number of unique hashtags u used	2.00e-01	2.04e-01	-2.57e+02	-2.91e+02

Table 2. Regression results of community-based and community-based-combining-context-based models

4 Conclusion

In this work, we employed the community detection, heterogeneous graph mining, and word2vec based textual similarity to investigate the effects of community-based and context-based features on information adoption and diffusion in Twitter. As Table 2 depicts, the usefulness of community-based features is verified; however, the context-based features are not shown to improve the prediction. Next, we will investigate other types of relations between users in Twitter network, and more sophisticated features will be utilized to characterize information adoption behavior on social media.

5 References

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*, 1–32.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World Wide Web* (pp. 591-600). ACM.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical review E, 80*(5), 056117.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREC* (Vol. 10, pp. 1320-1326).

Simos, G. C. (2015, August 19). How Much Data Is Generated Every Minute On Social Media? [Web log post]. Retrieved from http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012, April). 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). ACM.