Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Early Identification of Personalized Trending Topics in Microblogging

Liang Wu,[†] **Xia Hu,**[‡] **Huan Liu**[†]
[†]Computer Science and Engineering, Arizona State University, USA
[‡]Department of Computer Science and Engineering, Texas A&M University, USA
{wuliang, huanliu}@asu.edu, xiahu@tamu.edu

Abstract

Social media has become a primary platform for the spread of information. *Trending topics*, which are breaking news and immediately popular stories, have become an attractive data source facilitating the spread of emerging issues. Motivated by the diverse trending topics covering from sports to politics, it is essential to help users find personalized trending topics. Since a topic in social media may start trending and get obsoleted quickly, the personalization would be more valuable to a user if the trending topic can be recommended before it is outdated. In order to identify personalized trending topics at an early stage, we propose to identify and exploit the auxiliary information. In particular, through collectively modeling content of similar users with social network information, we identify additional past contents that can enrich the training data of trending topics and users. The key insight is that though *most* posts of a user may be irrelevant, a *few* key posts can be signals revealing interests towards a particular topic. Experiments on real-world data demonstrate that our proposed approach effectively personalizes trending topics when they just start trending.

Introduction

Microblogging has become a main platform for dissemination of emerging issues, and some news broke out on Twitter even before CNN. A recent study shows that 62% of American adults get news on social media[1]. Since various topics are trending simultaneously, it is critical to find a tailored list catering to users' interests. In this work, we aim to present a personalization system that tailors a personalized list of trending topics that are interesting to read for social media users.

A vital feature of trending topic personalization is its earliness. For example, the best timing to recommend topics for a baseball game is when it is ongoing since the stories become outdated soon after the match ends. Traditional approaches for personalization are incapable of dealing with trending topics since they rely on the accumulation of information, such as contents for content-based filtering, and user-item interactions for collaborative filtering. For trending topics, both kinds of data are generated with the topic going viral and becoming less attractive to read. Therefore, a key challenge of early personalization is to solve the cold-start problem.

Meanwhile, auxiliary information is pervasively present on social networks. An auxiliary data source is the historical posts of users. Figure 1 shows an example of user posts and Twitter trending topics on September 11^{th}, 2016. There were over 600 trending topics on Twitter that day in the United States, including "HillaryFaint" and "HillarysHealth" that were about Hillary Clinton's health issues, and "StanTheMan" which was about the US Open 2016 final. The preferences of the first user can be easily found because of the post. But for the second and third user, the interests can be easily found only if their past posts can be used, since the second user posted on men's single of US Open, and the third user was interested in Hillary's upcoming fundraising trips. Another auxiliary data source is the links between users. "Birds of a feather flock together", the principle of homophily reveals that friends on social networks are more likely to be interested in similar topics. A nice property of both kinds of auxiliary information is that they exist before a trending topic starts emerging, which can help solve the cold-start problem.

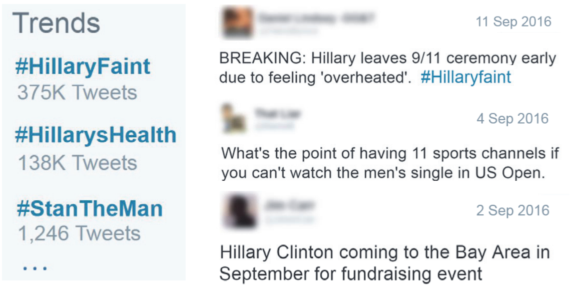
In this paper, we present a novel framework, Trending Topic Personalization approach (TTP), to personalize trending topics in an early stage. To solve the cold-start problem, TTP leverages social network structures to find the historical posts from like-minded users. To the best of our knowledge, this is the first work investigating personalization of trending topics on microblogging platforms. The main contributions of this paper are outlined as follows:

- We introduce the problem of personalizing trending topics in microblogging;
- Formulate a novel approach that jointly utilizes the network and additional content information;
- And conduct extensive experiments to validate the proposed model with real-world data.

Problem Statement

Let $\mathbf{U}$ denote the user set $\mathbf{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$. m represents the number of users and each user has a set of posts $\mathbf{u}_i =$

[1] http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/



Trends

#HillaryFaint
375K Tweets

#HillarysHealth
138K Tweets

#StanTheMan
1,246 Tweets

...

11 Sep 2016

BREAKING: Hillary leaves 9/11 ceremony early due to feeling 'overheated'. #Hillaryfaint

4 Sep 2016

What's the point of having 11 sports channels if you can't watch the men's single in US Open.

2 Sep 2016

Hillary Clinton coming to the Bay Area in September for fundraising event

Figure 1: Example user posts and trending topics on September 11^{th}, 2016. The first post explicitly includes the trending topic hashtag, and the later two are past posts of two users.

$\{\mathbf{p}_{i1}, \ldots, \mathbf{p}_{i|\mathbf{u}_i|}\}$. Each post is an attribute vector, *i.e.*, $\mathbf{p}_{ij} \in \mathbb{R}^n$, where n is the number of textual features. $\mathbf{y} \in \{-1, 1\}^m$ is the label vector denoting whether a user is interested in a topic. Given a trending topic, $\mathbf{y}_i = 1$ (user i is interested in the topic) if one of i's posts contains the hashtag of the trending topic, and $\mathbf{y}_i = -1$ otherwise. Let $\mathcal{A}$ denote the set of social links between microblogging users, where $a_{ij} = 1$ if i follows j and $a_{ij} = 0$ otherwise. We define the problem of personalizing trending topics as follows:

Given a trending topic, users $\mathbf{U}$*, the network information* $\mathcal{A}$*, and partial labels for training data* $\mathbf{y}$*, our goal is to learn an optimal function* f *that accurately predicts users in the test data who are interested in the topic.*

Personalizing Trending Topics

Content Modeling

Collaborative filtering models user interests by analyzing user-item correlations, which performs well when enough correlations are accumulated. However, a trending topic becomes popular immediately; so the correlations are not sufficient. Therefore, we aim to solve this problem by starting with a Content-Based Filtering (CBR) method. To predict a user's interests toward a trending topic based on content information, we adopt a logistic regression model, which has conventionally been used for CBR. The formulation of the optimization problem is shown as follows,

$$f(\mathbf{u}_i) = \frac{1}{1 + \exp(-\mathbf{w} \cdot \psi(\mathbf{u}_i) - b)}, \tag{1}$$

where $f(\mathbf{u}_i)$ denotes the prediction result that whether user i is interested in the trending topic. b is the model bias and $\mathbf{w}$ is the vector of model parameters. b and $\mathbf{w}$ are the parameters to optimize in a logistic regression model. $\psi(\cdot)$ maps a user to an attribute vector.

$\psi(\cdot)$ generates an attribute vector based on posts of users. For a user with a positive label ($\mathbf{y}_i = 1$), posts explicitly containing the trending topic are very few. Therefore, if only these posts are used, the corresponding attribute vector should be very sparse. If all posts of the user are selected, noisy information would be unavoidably included. An appealing model should be able to identify those implicitly correlated posts automatically. Motivated by the related research of computer vision, we propose to adopt Multi-Instance Learning (MIL) (Zhou 2004). In microblogging sites, a user contains a "bag" of posts and only few are related to a specific topic. We pose the personalization problem into an MIL task by reformulating Eq. (1),

$$f(\mathbf{p}_{ik}) = \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{p}_{ik} - b)}, \tag{2}$$

where $f(\mathbf{p}_{ik})$ predicts the label for a single post $\mathbf{p}_{ik}$. By aggregating prediction of all posts, the estimation of a user can be obtained as follows,

$$f(\mathbf{u}_i) = \frac{\sum_{k=1}^{|\mathbf{u}_i|} f(\mathbf{p}_{ik}) \cdot \exp(\alpha f(\mathbf{p}_{ik}))}{\sum_{k=1}^{|\mathbf{u}_i|} \exp(\alpha f(\mathbf{p}_{ik}))}, \tag{3}$$

where a softmax function is the aggregate results of a user. α is a parameter introduced to determine the extent of softness of the combination. Given the label vector $\mathbf{y}$, the optimal parameters $\mathbf{w}, b$ for a topic j can be obtained through minimizing the following cost function,

$$\epsilon(\mathbf{w}, b) = \frac{1}{2}\sum_{i=1}^{m}(y_i - f(\mathbf{u}_i))^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} + \gamma|\mathbf{w}|_1, \tag{4}$$

where $\mathbf{w}^T\mathbf{w}$ is a regularization term to avoid over-fitting by penalizing the model complexity, of which the extent is controlled by λ. Since among the many words only few are correlated with a trending topic, we invoke an ℓ_1 regularizer to induce sparsity.

According to the principle of homophily, friends are likely to have similar interests. We regard two users who follow each other as friends. Assume $\mathbf{E}$ represents friendship between users, where $\mathbf{e}_{it} = 1$ if $a_{it} = a_{ti} = 1$, and otherwise $\mathbf{e}_{it} = 0$. Therefore, the homophily can be modeled by minimizing the following regularizer

$$\sum_{e_{it} \in \mathbf{E}} e_{it}(f(\mathbf{u}_i) - f(\mathbf{u}_t))^2, \tag{5}$$

which smooths the prediction results of friends by penalizing the large difference between them. Motivated by graph learning literature, the regularizer can be rewritten as $f^t\mathcal{L}f$, where $f \in \mathbb{R}^m$ is the prediction results of users. $\mathcal{L}$ is the normalized *Laplacian* matrix of the corresponding social graph with the graph structure of $\mathbf{E}$. Specifically, the *Laplacian* $\mathbf{L}$ can be obtained through:

$$\mathbf{L} = \mathbf{D} - \mathbf{E},$$

where $\mathbf{D} \in \mathbb{R}^{m \times m}$ is a diagonal matrix and the diagonal elements are calculated as $d_{ii} = \sum_{k=1}^{m} e_{ik}$. The *normalized Laplacian* can then be calculated as:

$$\mathcal{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}.$$

Incorporating the normalized graph Laplacian norm as a regularizer rewrites the objective in Eq. (4) as follows:

$$\epsilon(\mathbf{w}, b) = \frac{1}{2}\sum_{i=1}^{m}(y_i - f(\mathbf{u}_i))^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} + \gamma|\mathbf{w}|_1 + \frac{\mu}{2}f^t\mathcal{L}f, \tag{6}$$

where the graph-based regularizer is reformulated and the resultant objective remains convex. μ controls the extent of penalization when the prediction results are different for friends. Since the amount of content information is massive, an efficient optimization method is required.

Model Fitting

For simplicity of presentation, we first augment $\mathbf{w}$ by incorporating b as $\mathbf{w}_0$, which can be implemented by adding an additional feature. Thus we aim to learn the optimal predictor as follows:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{m}(y_i - f(\mathbf{u}_i))^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} + \gamma|\mathbf{w}|_1 + \frac{\mu}{2}f^t\mathcal{L}f. \tag{7}$$

Since we employ a logistic model independent for each feature, features can be calculated separately when updating $\mathbf{w}$. Since both the normalized Laplacian and ℓ_1-regularizer are convex, we adopt projected gradient descent to update each feature w_k as follows:

$$\begin{aligned}\frac{\partial \epsilon}{\partial w_k} &= \sum_{i=1}^{m}(y_i - f(\mathbf{u}_i))\frac{\partial f(\mathbf{u}_i)}{\partial w_k} + \lambda w_k + \gamma \cdot \mathrm{Sign}(w_k) \\ &+ \mu \sum_{ij=1}^{m} \mathcal{L}_{ji} f(\mathbf{u}_i)\frac{\partial f(\mathbf{u}_i)}{\partial w_k}, \end{aligned} \tag{8}$$

where $\mathcal{L}_{ji}$ is the value of the corresponding entry in the normalized Laplacian matrix and $\frac{\partial f(\mathbf{u}_i)}{\partial w_k}$ is the gradient of softmax. The gradient of softmax can be further decomposed by each post $\mathbf{p}_{ij}$ as follows:

$$\frac{\partial f(\mathbf{u}_i)}{\partial w_k} = \sum_{j=1}^{|\mathbf{u}_i|}\frac{\partial f(\mathbf{u}_i)}{\partial f(\mathbf{p}_{ij})}\frac{\partial f(\mathbf{p}_{ij})}{\partial w_k}, \tag{9}$$

where the derivative of the logistic regression of posts can be computed by conventional approaches, and the derivative of the softmax aggregation function in terms of a post $\mathbf{p}_{ij}$ can be computed as follows:

$$\frac{\partial f(\mathbf{u}_i)}{\partial \mathbf{p}_{ij}} = \frac{(1 + \alpha f(\mathbf{p}_{ij}) - \alpha f(\mathbf{u}_i))\exp(\alpha f(\mathbf{p}_{ij}))}{\sum_{j=1}^{|\mathbf{u}_i|}\exp(\alpha f(\mathbf{p}_{ij}))}. \tag{10}$$

Content-centric Features

We derive four types of content features in this work. **Words** directly characterize the content of posts. **Hashtags** are indicative for semantic of a post. We remove hashtags of trending topics. The **bigram** features of hashtags and words can represent semantics. Also, we use the **sentiment** polarity of sentiment words, phrases and emoticons in posts as features. We use the description[2] to derive sentiment for emoticons and use SlangSD for words and phrases (Wu, Morstatter, and Liu 2016).

Experiments

Experimental Settings

To collect for the dataset, we randomly collect 1,012 trending topics from Twitter, from June 6^{th}, 2016 to June 8^{th}, 2016 in the area of United States. In order to find potentially interested users, we randomly collect users who post during that period from Twitter's Streaming API[3]. For each user, we obtain their followers and friends to build up the adjacency matrix and collect up to 20 most recent posts. Statistics on the dataset are shown in Table 1. We use the posts that are generated within the first hour when the topic starts trending as training data, and we use users who post on the trending topic after the first hour for test. We filter out users that are likely to be content polluters (Wu et al. 2017a; 2017b).

Table 1: Statistics of the dataset used in this study.

Topics	Users	Labeled Posts
1,012	101,351	10,151
Posts	Links	
2,015,802	20,046,715	

There are three positive parameters involved in the experiments, including λ, γ, and μ in Eq.(7). As a common practice, all the parameters can be tuned through cross-validation. We set $\lambda = 0.1$, $\gamma = 0.1$, and $\mu = 0.1$, though in our experience the parameters do not significantly impact performance. We follow standard personalization settings to evaluate the performance and adopt the metric of "root-mean-square-root" (RMSE).

Performance Comparisons

We compare with the following state-of-the-art personalization methods:

- *PMF*: Collaborative filtering has been regarded as a state-of-the-art recommendation method in various areas such as movies. In this work, we adopt BPMF (Salakhutdinov and Mnih 2008), which adopts fully Bayesian treatment of the Probabilistic Matrix Factorization.
- *CBF* and *CBF+*: We include two Content-Based Filtering methods. In CBF, we use posts hashtagged with the trending topic to represent a user. While in CBF+, all posts of an interested user are used.
- *LMGR* and *LMGR+*: LMGR jointly utilizes content and network information (Zhang, Popescul, and Dom 2006). Similarly, LMGR exploits only the posts hashtagged by the trending topic while LMGR+ uses all posts.
- *SocDim*: SocDim (Tang and Liu 2009), which learns user interests by projecting social relations into a low dimensional space, and has commonly been used for categorizing users. It can be considered as a state-of-the-art method for relational learning on social networks.
- *Random*: Because there are much more negative training examples than the positive examples in the dataset, the absolute value of RMSE is not very meaningful. For comparison purposes, we also use a Random baseline that uniformly selects trending topics for each user.

[2] http://emojipedia.org/twitter/

[3] https://dev.twitter.com/streaming/reference/post/statuses/filter

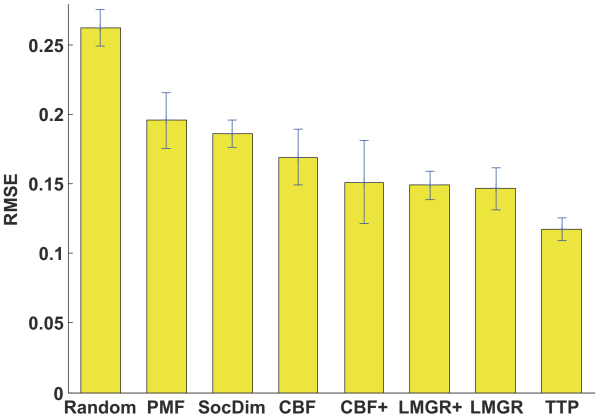


Figure 2: Performance comparisons for different personalization methods.

The comparison is shown in Figure 2. Since there are much more negative training data instances than the positive ones in the Twitter dataset, the absolute value of RMSE is not very meaningful. The RMSE of Random baseline is 0.2621. We are expecting the relative decreasing of RMSE for a more effective approach. Based on the results, we draw following observations. Traditional recommendation approaches, *i.e.*, content-based (CBF, CBF+) and collaborative filtering (PMF) cannot effectively personalize trending topics. SocDim outperforms PMF and Random, showing that knowledge that is useful for identifying user interests exists in the network structures. The proposed TTP outperforms existing methods that directly integrate social network structures with user contents (LMGR, LMGR+) by selecting the related content from the massive amount of historical information, instead of ignoring or adopting all. The result demonstrates that TTP is effective in inspecting user interests and personalizing trending topics.

Earliness of Personalization

A key objective of our study is to find interesting trending topics at an early stage before they become obsolete. More user posts and other data are available for training at a later stage with the topic being trending. However, late recommendations are much less practically useful, since a topic trended yesterday may get outdated and less interesting to read. Therefore, we investigate how effective TTP is when less training data is available during the early period.

The results are shown in Figure 3. In order to evaluate earliness, we train models by additionally using training data by its chronological order. In particular, we additionally add training data based on the time order they were generated after the trending topic started trending. Since SocDim only uses the network information, the performance is constant. The RMSE of other methods decreases with more training data being added. According to the results, the best baseline, LMGR, achieves RMSE of TTP with training data of first ten minutes by a lag of 90 minutes. Therefore, the empirical results show that the use of TTP not only yields low error rate, but also finds interesting trending topics hours faster than traditional personalization approaches.

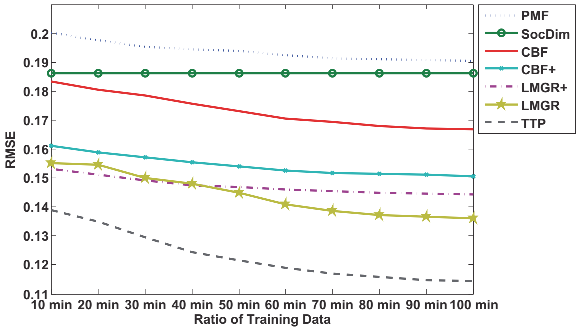


Figure 3: Performance of different models with chronologically additional training data.

Conclusion

In this work, we propose the Trending Topic Personalization approach (TTP) to personalize trending topics at an early stage. TTP tackles the cold-start problem through jointly exploiting social media posts and social interactions. Through experiments on real-world data, we have demonstrated the gains of performance and earliness of the proposed method.

Acknowledgments

The work is funded, in part, by ONR N00014-16-1-2257 and the Department of Defense under the MINERVA initiative through the ONR N000141310835.

References

Salakhutdinov, R., and Mnih, A. 2008. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *International Conference on Machine learning*, 880–887. ACM.

Tang, L., and Liu, H. 2009. Relational learning via latent social dimensions. In *ACM International Conference on Knowledge Discovery and Data Mining*.

Wu, L.; Hu, X.; Morstatter, F.; and Liu, H. 2017a. Adaptive spammer detection with sparse group modeling. In *ICWSM*.

Wu, L.; Hu, X.; Morstatter, F.; and Liu, H. 2017b. Detecting camouflaged content polluters. In *ICWSM*.

Wu, L.; Morstatter, F.; and Liu, H. 2016. Slangsd: Building and using a sentiment dictionary of slang words for short-text sentiment classification. *arXiv preprint arXiv:1608.05129*.

Zhang, T.; Popescul, A.; and Dom, B. 2006. Linear prediction models with graph regularization for web-page categorization. In *International Conference on Knowledge Discovery and Data mining*, 821–826. ACM.

Zhou, Z.-H. 2004. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*.